Eliciting Bootstrapping: The Development of Introductory Statistics Students’ Informal Inferential Reasoning via Resampling

Jeffrey Allen McLean
Syracuse University

Follow this and additional works at: https://surface.syr.edu/etd

Part of the Education Commons

Recommended Citation
https://surface.syr.edu/etd/396

This Dissertation is brought to you for free and open access by the SURFACE at SURFACE. It has been accepted for inclusion in Dissertations - ALL by an authorized administrator of SURFACE. For more information, please contact surface@syr.edu.
Abstract

Bootstrapping has become an important tool for statisticians, who assert that it is intuitive to novice statistics students. The process of collecting bootstrap resamples was once time demanding, but technology now allows data collection to be performed nearly instantaneously. This study focuses on the construction and development of secondary and tertiary introductory statistics students' (n=68) reasoning about bootstrapping and informal inference. Students engaged in a four-week instructional unit designed as two model development sequences. Through the use of hands-on manipulatives and technology, students constructed and developed reasoning of sampling, the resampling process of bootstrapping, and inference. The focus of my analysis was on the model development of four focus groups of students. Groups of students constructed models of sampling and inference that they used to collect samples, aggregate the samples to form an empirical sampling distribution, and use the aspects of this distribution of samples to make claims about the population from which the samples were drawn. I summarize and categorize groups of students’ models and trace the development of the focus groups’ models throughout the unit.

Simulation of data led some students to develop a global view of the randomness of sampling and reason with multiple aspects of empirical sampling distributions to draw inferential claims. Some students applied a multiplicative view of the sample and global view of the sampling process to construct a resampling process similar to bootstrapping, but fell short of constructing the bootstrapping process by not collecting resamples that were equal in size to the original sample. Class discussion of a follow-up activity, similar in structure to the model eliciting activity, encouraged students to consider the value of drawing resamples of equal size to the population and led students to construct the method of bootstrapping. Students then used the
bootstrapping process to drawn inferences about one population and to compare two populations of data.

The findings of this study contribute to the field of statistics education by examining student thinking while constructing and developing bootstrapping methods, as well as investigating the relationship between this thinking and the drawing of informal inferences. This study demonstrates that it is possible for model development sequences to elicit and develop students’ models of bootstrapping. With the trend in statistics education of moving from the focus on theoretical distributions towards the simulation and analysis of data, these findings have implications towards the design of future introductory statistics curricula.
ELICITING BOOTSTRAPPING: THE DEVELOPMENT OF INTRODUCTORY STATISTICS
STUDENTS’ INFORMAL INFERENTIAL REASONING VIA RESAMPLING

By

Jeffrey A. McLean

B.S., Applied Mathematics, Rochester Institute of Technology, 2005
M.S., Mathematics, University of Connecticut, 2009

Dissertation
Submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Mathematics Education

December 2015
# Table of Contents

Chapter 1 – Introduction .......................................................................................... 1
  Aims of Research ................................................................................................. 6
  Repeated Sampling ............................................................................................. 7
  Resampling ........................................................................................................... 8
  Research Questions ............................................................................................ 11

Chapter 2 – Literature Review ................................................................................. 12
  Informal Inferential Reasoning ........................................................................... 12
  Technology Use in Statistics Education ............................................................. 27
  Modeling .............................................................................................................. 30
  Summary ............................................................................................................. 33

Chapter 3 – Design and Methodology ................................................................... 35
  Research Design ................................................................................................. 35
  Settings and Participants ..................................................................................... 36
  Model Development Sequences ........................................................................ 37
  Data Collection .................................................................................................. 48
  Data Analysis .................................................................................................... 51

Chapter 4 – Results ................................................................................................ 55
  Pre/Post-Test Results ........................................................................................ 57
  Models of Sampling and Inference .................................................................... 58
  Models of Resampling and Inference ................................................................. 111
  Summary of Model Development .................................................................... 157
List of Tables and Figures

Tables

Table 1. Intended models in the first model development sequence ............................................. 39
Table 2. Intended models in the second model development sequence ........................................ 46
Table 3. Focus group participants ............................................................................................... 51
Table 4. Summary statistics and one-way ANOVA analysis of the pre-test scores between the two settings ......................................................................................................................... 57
Table 5. Summary statistics and paired t-test for increase in pre/post-test scores ..................... 58
Table 6. Models of resampling...................................................................................................... 114

Figures

Figure 1. Visual representation of the process of repeated sampling........................................ 8
Figure 2. Visual representation of the process of resampling using bootstrapping methods.. 10
Figure 3. Concept of distribution ............................................................................................. 15
Figure 4. Overview of the activities in the two model development sequences ..................... 38
Figure 5. TinkerPlots environment for the Ophelia MXA ...................................................... 41
Figure 6. Bin of multi-colored beads for the Nike sneakers MXA and MAAs. ..................... 43
Figure 7. TinkerPlots environment for the Nike sneakers MXA .............................................. 44
Figure 8. Craft sticks used for the Mixed Nuts MEA .............................................................. 47
Figure 9. TinkerPlots environment for the Mixed Nuts MEA ............................................... 48
Figure 10. Visual representation of focus group D’s collected samples for Ophelia’s predictions ............................................................................................................................................. 63
Figure 11. Non-focus groups’ table for the probabilities of guessing the numbers of correct predictions ........................................................................................................................................... 64
Figure 12. Non-focus group’s tree diagram for Ophelia’s eight predictions .................. 66
Figure 13. Megan’s solution to Paul the Octopus Controversy. .............................. 68
Figure 14. Focus group A’s dotplots for each student’s 10 sets of Ophelia’s predictions...... 71
Figure 15. Henry’s dotplot for 50 trials of Ophelia’s predictions (focus group B)........... 74
Figure 16. George’s dotplot for 50 trials of Ophelia’s predictions (focus group B)......... 75
Figure 17. Non-focus group student’s table for the number of samples collected for numbers of correct predictions. ................................................................. 78
Figure 18. Focus group D’s dotplot of Nike sneaker purchases .................................. 80
Figure 19. Focus group C’s dotplot of Nike sneaker purchases ................................. 83
Figure 20. Aggregate data of 105 samples of 20 sneaker purchases. .......................... 82
Figure 21. Focus group A’s dotplot of Nike sneaker purchases ................................. 85
Figure 22. Henry’s (focus group B) dotplot for the number of Nike sneaker purchases in samples of 20 after collecting 120 samples. .................................................. 91
Figure 23. Lena’s (focus group C) sketch for the number of Nike sneaker purchases for 1000 samples ......................................................................................... 92
Figure 24. Veronica’s (non-focus group) dotplot for the number of Nikes in samples of 20 sneaker purchases for 55 samples ......................................................... 94
Figure 25. Veronica’s (non-focus group) predicted dotplot for the number of Nikes in samples of 20 sneaker purchases after collecting an additional 114 samples .......... 94
Figure 26. Focus group D’s samples of Nike and Adidas sneaker purchases ................. 99
Figure 27. Non-focus group’s sketch of the anticipated spread of samples for the difference between Nike and Adidas sneaker purchases ........................................... 102
Figure 28. Non-focus group’s dotplot for the number of Nikes in 20 samples of 20 sneaker purchases.......................................................... 104

Figure 29. Non-focus group’s dotplot for the number of Adidas in 20 samples of 20 sneaker purchases.......................................................... 104

Figure 30. Focus group A’s table of how many more Nikes were in their samples compared to Adidas sneakers ................................................................. 106

Figure 31. Focus group A’s dotplot of how many more Nikes were in their samples compared to Adidas sneakers ................................................................. 107

Figure 32. Focus group A’s table and dotplot for the counts of peanuts in 10 resamples of mixed nuts, collected without replacement...............................116

Figure 33. Focus group C’s table for the count and percentage of peanuts in five resamples of mixed nuts, collected without replacement...............................119

Figure 34. Focus group C’s likely range for the percentage of peanuts in the brand of mixed nuts ..........................................................................................119

Figure 35. Focus group C’s dotplot for the counts of peanuts in 7 resamples of 5 mixed nuts, collected without replacement ..............................................120

Figure 36. Non-focus group’s dotplot for the count of peanuts in 14 resamples of 10 mixed nuts, collected without replacement.................................121

Figure 37. Non-focus group’s table and dotplot for 20 resamples of 10 mixed nuts, collected without replacement.................................................................121

Figure 38. Focus group C’s table and dotplot for the counts of peanuts in 17 resamples of 10 mixed, collected without replacement ........................................122
Figure 39. Yon’s non-focus group’s table for the number of peanuts and other nuts, collected without replacement, with varying resample sizes ................................................. 124

Figure 40. Non-focus group data and dotplot for the percentage of peanuts in 10 resamples of varying sizes, collected without replacement ......................................................... 126

Figure 41. Non-focus group’s data for counting peanuts and other nuts in two resamples of 20 mixed nuts, collected with replacement ................................................................. 127

Figure 42. Spinner constructed by a non-focus group to resample the mixed nuts .......... 133

Figure 43. Non-focus group’s Mixed Nuts MEA approach and their data ....................... 134

Figure 44. TinkerPlots environment for Trident Gum projected in the front of the classroom collecting 1000 samples of five dentists ................................................................. 134

Figure 45. Aggregated class dotplot for bootstrap samples of mixed nuts collected with hands-on manipulatives ................................................................. 144

Figure 46. Lena’s (focus group C) dotplots of simulated data for the number of peanuts in bootstrap samples of 25 mixed nuts for each brand of mixed nuts ................. 152

Figure 47. Nate’s (focus group C) combined dotplot for simulated data for the number of peanuts in bootstrap samples of 25 mixed nuts for each brand of mixed nuts. .. 153
Chapter One - Introduction

Over the past few decades, statistics education has become an integral part of the mathematics curriculum at all levels. Statistics education gained traction in the K-12 classroom with the release of the National Council of Teachers of Mathematics' (NCTM) influential documents *Curriculum and Evaluation Standards for School Mathematics* (NCTM, 1989) and its successor, *Principles and Standards for School Mathematics* (NCTM, 2000). Data analysis and probability was highlighted as one of the five strands of mathematical content that should be emphasized in K-12 education. This placed the learning of statistics on equal footing with number sense, algebra, geometry, and measurement in the recommended school curriculum.

Prior to these documents, statistics was often “the mere frosting on any mathematics program if there was time at the end of the school year” (Shaughnessy, 2007, p. 957). The *Common Core State Standards for Mathematics* (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010) also emphasize statistics and probability as a domain of standards for students in the 6th through 12th grades, with statistical topics covered in the domain of measurement and data for grades K-5. This lesser emphasis on topics in statistics at the early grade levels is a noted critique of these standards (Rossman & Shaughnessy, 2013).

In addition to these standards documents citing the need for an emphasis on statistics education at the K-12 level, the *Guidelines for Assessment and Instruction in Statistics Education* (GAISE) College Report (Aliaga et al., 2005), funded by the American Statistical Association (ASA), has noted a growth in statistics education at the tertiary level. The report cites that in the past, statistics was often one course, taught to a narrow group of future scientists, while today, it has transformed into a family of courses that are taught to many students with varying interests and goals.
The GAISE PreK-12 Report asserts that the ultimate goal of statistics education is statistical literacy (Franklin et al., 2007), which is defined as the ability “to use sound statistical reasoning to intelligently cope with the requirements of citizenship, employment, and family and to be prepared for a healthy, happy, and productive life” (p. 1). Those who are statistically literate possess statistical skills that allow them to take advantage of data to be well informed or make good decisions (Shaughnessy, 2007). The GAISE PreK-12 Report cites examples of statistical literacy as: the understanding of public opinion polls, making personal choices such as using statistical information regarding the safety and effectiveness of medications, the role of statistics in fields such as medicine to exhibit scientific progress, and skepticism of statistical findings. Konold and Higgins (2003) took a strong stance towards the importance of statistical literacy and asserted that:

At the practical level, knowledge of statistics is a fundamental tool in many careers, and without an understanding of how samples are taken and how data are analyzed and communicated, one cannot effectively participate in most of today’s important political debates about the environment, health care, quality of education, and equity. For those who have traditionally been left out of the political process, probably no skill is more important to acquire in the battle for equity than statistical literacy. (p. 193)

Statistical literacy goes beyond being the understanding required for specific career paths, and encompasses understanding that is deemed to be essential for participating in today's society.

In addition to promoting statistical literacy, the GAISE College Report (Aliaga et al., 2005) asserted that students need to develop statistical thinking. While statistical literacy involves understanding the language and fundamental ideas of statistics, statistical thinking moves on from understanding how others have used statistics, to acquiring the reasoning used by
statisticians when approaching statistical problems. Those exhibiting statistical thinking understand the need for data, the importance of data production, the omnipresence of variability, and the quantification and explanation of variability (Cobb, 1992). To investigate the structure of statistical thinking, Wild and Pfannkuch (1999) developed a four dimensional model, categorizing these dimensions as the investigative cycle, types of thinking, the interrogative cycle, and dispositions. The investigative cycle describes how one acts during the process of a statistical investigation. A cycle in this process contains the formation of a problem, a plan to understand the problem, the collection of data, the analysis of the data, and the reaching of conclusions. The second dimension, types of thinking, describes both general types of thinking and those fundamental to statistical thinking. Examples of general types of thinking described are the planning for and anticipation of problems, and the use of problem solving skills. Those types of thinking fundamental to statistics are the consideration for variation and transnumeration. Wild and Pfannkuch define transnumeration as “numeracy transformations made to facilitate understanding” (p. 227). This concept relates to the dimensions of statistical investigation such as determining what forms of data are meaningful for an investigation and changing the representations of graphical data in order to show new meaning.

The third dimension, the interrogative cycle, is a general thinking process, not unique to statistics, which is used throughout a statistical investigation. The cycle begins with brainstorming ideas and continues with seeking out information relating to these ideas. The cycle continues with interpreting, criticizing, and judging these ideas. Judgment may then lead to a new cycle beginning with the generation of new ideas. The final dimension of statistical thinking discusses dispositions vital to possess during statistical investigation, such as imagination, skepticism, and a propensity to seek for a deeper meaning.
Statistical literacy and thinking are broad goals of statistics education. I focused on one foundational concept within these goals, the making of inferences about a population based on sample data. This concept has been highlighted by standards and policy documents (Aliaga et al., 2005; National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010; NCTM, 2000). In upper secondary and introductory college level coursework, inferences come in the form of formal inference, which generally revolves around the use of theoretical probability distributions to create confidence intervals or perform hypothesis testing. In a traditional statistics classroom, this involves the use of textbook problems that generally require only the use of given procedures. Students may become proficient at applying the given procedures without possessing the understanding of why the given procedure is used to make formal inferential claims (Aliaga et al., 2005). Students should understand the meaning of their answers, rather than only have the ability to make the calculations (Aliaga et al., 2005).

Research has indicated that informal inferential reasoning may support the development of students’ understanding of formal inferential reasoning (Bakker & Gravemeijer, 2004; Makar & Rubin, 2009; Pfannkuch, 2005; Saldanha & Thompson, 2002). Informal inferential reasoning has been defined as “the drawing of conclusions from data that is based mainly on looking at, comparing, and reasoning from distributions of data” (Pfannkuch, 2007, p. 149) and “the way in which students use their informal statistical knowledge to make arguments to support inferences about unknown populations based on observed samples” (Zieffler, Garfield, delMas, & Reading, 2008, p. 44). Informal inference highlights some of the mechanics behind formal inferential procedures and the reasoning behind inferential claims. This is in line with policy documents such as the GAISE College Report (Aliaga et al., 2005), which recommends that introductory statistics courses not merely discuss the procedures, but delve into the concepts behind them.
A current trend in statistics education is the shift from a focus on theoretical distributions and numerical approximations into an emphasis on data production and analysis (Cobb, 2007). Cobb asserts that many statistics curricula are outdated and based on how statistics could be learned prior to the computing power of modern times. There need to be new statistics curricula that are focused on inference, rather than focused on the normal distribution. Cobb continued by stating that:

...the consensus curriculum is still an unwitting prisoner of history. What we teach is largely the technical machinery of numerical approximations based on the normal distribution and its many subsidiary cogs. This machinery was once necessary, because the conceptually simpler alternative based on permutations was computationally beyond our reach. Before computers statisticians had no choice. These days we have no excuse. Randomization-based inference makes a direct connection between data production and the logic of inference that deserves to be at the core of every introductory course. (2007, p. 1)

Decades ago, the use of approximation techniques were needed due to the tediousness of collecting large amounts of samples, but technology is now capable of collecting many samples nearly instantaneously. This advance in technology should have an impact on the statistics curriculum since it allows the use of data simulation to illustrate the logic of formal inference.

Cobb further asserted that, “There’s a vital logical connection between randomized data production and inference, but it gets smothered by the heavy, sopping wet blanket of normal-based approximations to sampling distributions” (2007, p. 7). For constructing the theoretical sampling distributions for various situations, such as the mean, the difference between two means, and the difference between two proportions, there are different procedures and formulas
that are needed. To draw inferences from the normal approximation of the sampling distribution, students must understand the relationship between this theoretical distribution and the drawing of samples. In the case of randomized data production, numerical approximations are taken out of the procedure, which were only there out of a dated necessity.

Aims of Research

With Cobb’s assertions in mind, I focused in this study on the development of students’ models of the process of bootstrapping while engaging in data driven sampling activities in two model development sequences. Through the use of hands-on manipulatives and simulations with technology, the activities elicited the development of participants’ models of sampling and the inferences that could be drawn from empirical sampling distributions. These models were then adapted to elicit the process of bootstrapping. While limited research has been done on student learning of statistics with bootstrapping methods (Garfield, delMas, Zieffler, 2012; Pfannkuch & Budgett, 2014; Pfannkuch, Forbes, Harraway, Budgett, & Wild, 2013), researchers have asserted that bootstrapping may promote student learning of the logic of inference (Cobb, 2007; Engel, 2010; Hesterberg, 2006)).

The bootstrapping method has already become part of introductory statistics coursework, with textbooks such as *Statistics: Unlocking the Power of Data* (Lock, Lock, Lock-Morgan, Lock, & Lock, 2013) introducing the method of bootstrapping to define confidence intervals well before discussing confidence intervals with normal approximation methods. Lock et al. claim that the bootstrapping method has become an important tool for statisticians and is also very intuitive to novice statistics students and accessible at the early stages of statistics coursework. They further state that bootstrapping capitalizes on students' visual learning skills and helps to build students' conceptual understanding of key statistics idea. In this study, I investigated these
claims by examining how students develop the bootstrapping process and use it to make informal inferential claims. To develop bootstrapping, I used two types of sampling activities, repeated sampling and resampling.

**Repeated Sampling**

Sampling activities, such as those recommended by Saldanha and Thompson (2002), often involve repeated sampling from a known population. I refer to repeated sampling activities as those where many samples are drawn from a population. The process of repeated sampling can be thought of as starting with the population distribution and taking a random sample of \( n \) elements, with these \( n \) elements having a sample distribution. A statistic, such as a proportion or mean, is then calculated from the sample distribution, which then becomes an element of the empirical sampling distribution. Returning to the population, this cycle is continued many times in order to construct the empirical sampling distribution (Figure 1). Many samples are drawn from the population in order to approximate the sampling distribution of all possible sample statistics. From this distribution of statistics of samples, the mean value of the sampling distribution approximates a population parameter, such as the mean.
Like the process of bootstrapping, repeated sampling activities can take advantage of modern computing power to construct an empirical sampling distribution. An issue with repeated sampling activities is that while they allow students to learn how to construct and make statistical claims from empirical sampling distributions, it is not always practical to collect many samples from a population. In practice, statisticians likely only have one sample from which to draw their conclusions. Processes of resampling, such as bootstrapping, may be a possible bridge between how the sampling distribution is constructed, the inferences made from the drawing of many samples, and the similar distribution and inferences made from drawing only one sample.

**Resampling**

Resampling activities, just as in repeated sampling, work towards constructing an empirical sampling distribution. Both forms of activities draw many samples, but the difference is in how these samples are drawn. In repeated sampling many samples are drawn from a population, while in resampling, only one sample is drawn from the population. Depending on
the method of resampling, this one sample is then used to construct many additional samples, often called resamples, which can then be used to construct an empirical sampling distribution. The form of resampling that this study investigated was bootstrapping. Efron (1979) introduced the method of bootstrapping as an alternative to earlier resampling methods. He asserted that the bootstrap was more widely applicable and dependable than earlier resampling methods, while also using a simpler procedure. The term bootstrap is used in the idiom to pull oneself up by one's bootstraps, which means to accomplish a goal with the resources on hand. Rather than take many samples to estimate a sampling distribution, the bootstrap method uses the one sample on hand to accomplish the same goal.

Bootstrapping begins with one sample of data from the population, which is assumed to represent an approximation of the population's distribution. From this initial sample, bootstrap samples (also called resamples) are constructed by choosing \( n \) elements, \textit{with replacement}, creating many equal sized samples. A statistic from each of these bootstrap samples is then calculated and aggregated to form an empirical bootstrap sampling distribution (Figure 2). Since the original sample from the population was used to approximate the population's distribution, the assumption is that the data values in the sample represent the distribution of all data values in the population, and the proportion of each data value in the sample represents the proportion of that data value in the population. If done with hands-on manipulatives, this sampling process using replacement can be more cumbersome and time demanding than the process of repeated sampling, but technology can be used to simulate this procedure in a short period of time.
While the concept of using one sample, rather than many, to make inferential claims may bridge the gap from repeated sampling to formal inference, the concept of bootstrapping is complex. It comes with the assumption that the original sample is typical of any sample taken from the population. Resampling methods, such as bootstrapping, could provide a means for students to use their understanding of the empirical sampling distribution and repeated sampling when the population's distribution is not known. Because of this, this study first focused on students' concepts of the empirical sampling distribution through repeated sampling, and then how they transition these understandings into the concept of resampling. Sampling activities, especially resampling, can be tedious and require much work to generate data. Technology is crucial to allow students to see the results of generating much more data than is feasible with hands-on manipulatives. I investigated how the manipulatives may be crucial in students'
understanding of the processes behind sampling, which may be hidden in simulations with technology.

**Research Questions**

To investigate students' learning of bootstrapping methods, my study was guided by the following questions:

1) What student reasoning develops as students move from repeated sampling methods to bootstrapping methods?

2) How do students develop their reasoning about bootstrapping methods in order to make informal inferences about a population of data?

3) What student reasoning is revealed and supported by moving from using hands-on manipulatives to computer simulations during repeated sampling and bootstrapping activities?
Chapter Two - Literature Review

This study focuses on students' development of an understanding of inferential reasoning through the implementation of an instructional unit constructed as two model development sequences centered on repeated sampling and resampling activities. I begin by discussing how informal inferential reasoning has been defined by researchers, including the role of distributional reasoning in informal inferential reasoning. I then discuss informal inferential reasoning within the context of repeated sampling and resampling activities. Since this study used both hands-on manipulatives and computer simulation methods as part of the instructional unit, I then discuss the role of technology in statistics education, and specifically informal inferential reasoning. Lastly, the design of this study and methods of data analysis are influenced by the use of modeling and model development sequences, so I discuss research in this area and how it influenced this study.

Informal Inferential Reasoning

Informal inferential reasoning has been defined as “the drawing of conclusions from data that is based mainly on looking at, comparing, and reasoning from distributions of data” (Pfannkuch, 2007, p. 149), “the process of making probabilistic generalizations from (evidenced with) data that extend beyond the data collected” (Makar & Rubin, 2007, p.1), and “the way in which students use their informal statistical knowledge to make arguments to support inferences about unknown populations based on observed samples” (Zieffler, Garfield, delMas, & Reading, 2008, p. 44). In the latter definition, informal statistical knowledge refers to inferring about the possible characteristics of a population or the possible differences in characteristics between two populations based on observations of characteristics such as shape and center in sample(s). The fundamental difference between descriptive statistics and inferential statistics is the act of trying
to form meaning beyond the sampled data. Researchers have shown that learners of statistics have difficulty understanding the differences between discussing meaning from a sample versus a population (Chance, delMas, & Garfield, 2004; Makar, & Rubin, 2009; Pfannkuch, 2007; Saldanha & Thompson, 2002).

Makar and Rubin (2009) posited three principles that they think are key to informal inferential reasoning: 1) the making of generalizations beyond describing the given data; 2) the use of data as evidence for those generalizations; and 3) the use of probabilistic language to describe the generalizations. The first principle asserts the importance of the relationship between sample data and the population from which the sample is drawn. Key to informal inferential reasoning is the leap from descriptive statistics, which draws meaning about the sample data, to inferential statistics, which may then use the descriptive statistics to draw conclusions about the larger population. The second principle states that learners should use data as evidence for arguments or explanations. These data could be in any form, such as descriptive or observational, and appropriate for the community in which the data are presented. Makar and Rubin use the word community to emphasize that learners at different ages and understandings of informal inferential reasoning may use different forms of data as their evidence. For instance, younger students may be more prone to use observational data as evidence, but as their understanding progresses, they may use more robust forms of data, such as descriptive data. The final principle of using probabilistic language relies on the idea that since the generalizations being made go beyond the data, the data cannot support the generalizations in absolute terms. Makar and Rubin were not suggesting the use of quantifiable levels of confidence with this principle, but an informal view of the strength of predictions, arguments, or explanations based on the data used to construct them. An example of this principle would be using the data to
predict a likely range of values, rather than asserting a specific value is the likely outcome.

Learners must have an understanding of how to interpret the distribution of a sample or samples of data in order to make claims about future samples or populations of data with some uncertainty. I first discuss the research on distributional reasoning where learners interpret samples of data and then how learners use distributional reasoning to make informal inferences.

**Distributional reasoning.** In statistics, we are rarely interested in viewing a data set as a group of individual values. Conclusions are more likely to be drawn from the discernible patterns of these individual data points (Wild, 2006). “Until a data set can be thought of as a unit, not simply as a series of values, it cannot be described and summarized as something that is more than the sum of its parts” (Mokros & Russell, 1995, p. 35). Calculated measures from a data set such as the mean, standard deviation, and range are important for describing a distribution of data, but it is only when the many aspects of a distribution are viewed together can one comprehend the larger picture of the data. To elaborate on this larger picture, I discuss frameworks describing what it means to exhibit distributional reasoning.

Bakker and Gravemeijer (2004) discussed a framework for the concept of distribution by viewing the relationship between a distribution, as a conceptual entity, and its data, as individual values. Aspects of the distribution for data sets include center (mean, median, and mode), spread (range, standard deviation, and interquartile range), density (relative frequency, majority, and quartiles), and skewness (position of majority of data). The relationship between these aspects of distribution is shown in Figure 3. Bakker and Gravemeijer asserted that it is typical for novices in statistics to first focus on a view of data as individual values that can be used to calculate each aspect, such as the center. From this focus on data, the novice may calculate measures of center, but not have a notion of the mean, median, or mode as a representative measure of center (Groth
& Bergner, 2006). Experts in statistics have the understanding to view the data as individual values to perform various calculations, but also a top down view, where the distribution of the data as a whole is first considered to determine an understanding of the meaning of the aspects in relation to all data values.

![Figure 3. Concept of distribution (Bakker & Gravemeijer, 2004, p. 148)](image)

Friel, O’Connor, and Mamer (2006) stated a similar view of distribution as Bakker and Gravemeijer (2004) and discussed student understanding of this concept in an analogous manner. Friel, O’Connor, and Mamer discussed a process for how students do statistics when exploring a problem using statistics and data analysis. The process is made up of four phases: posing question(s), collecting data, analyzing the distribution(s), and interpreting the results. During the phase of analyzing the distribution(s), Friel et al. posited three levels of observations that students may show as developing knowledge of the conception of distribution. In the first level, students view data as individual points without seeing how the values might be related. In the second level, students identify groups of the data with similar characteristics, such as categories or clusters. In the third level, students begin to view the data values in aggregate as an object and focus on characteristics of the distribution as a whole. The first and third levels are similar to the bottom-up and top-down view of the concept of distribution (Bakker & Gravemeijer, 2004). Friel, O’Connor, and Mamer's framework emphasizes the hierarchy of these phases while
detailing the second level as a path between the first and third levels.

Reading and Reid (2006) also proposed a framework related to how learners develop knowledge of distribution. The authors began with the concrete-symbolic mode of the SOLO (Structure of the Observed Learning Outcome) taxonomy (Biggs & Collis, 1982) as the basis of their framework. The four levels within this mode are: prestructural, which shows no focus of relevant aspects of a concept; unistructural, which focuses on one aspect of a concept; multistructural, which shows focus on multiple unrelated aspects of a concept; and relational, which focuses on several aspects of a concept and the relationships between these aspects.

Reading and Reid chose to consider five relevant aspects of distribution: center, spread, density, skewness, and outliers. The first four aspects are the same as those used by Bakker and Gravemeijer (2004) in their distribution framework. Reading and Reid assert that deviations from the general pattern of the distribution are an important attribute and include outliers as a separate aspect.

With the SOLO framework and these aspects of distribution Reading and Reid (2006) created two cycles of levels of development of reasoning about and with distributions. The first cycle pertains to the understanding of key elements of a distribution while the second discusses the reasoning of using a distribution for statistical inference. Within the first cycle, learners showing prestructural reasoning do not refer to any of the five key aspects of a distribution. The authors posit that this may be due to learners' problems understanding numerical or graphical representations. Those showing unistructural reasoning refer to only one key aspect of distribution. For example, a learner in this level may compare two distributions solely based on their mean values without taking other elements such as variation into account. Multistructural reasoning shows understanding of more than one of the five aspects of distribution, but does not
display the links between these aspects. Learners at this level may discuss the shape of a
distribution using standard terminology such as skewness and outliers or nonstandard language
such as clumped and compact in the bottom 50% of the data values. These learners exhibit
reasoning of the existence of these aspects of a distribution but may not take them into account
when working with the distribution. Reading and Reid give an example where learners chose to
ignore the existence of outliers and use the smaller data set for inference without taking into
account the effect that the outliers may have on their decisions. In the final level, relational,
learners show reasoning with key elements of a distribution, such as the link between center,
spread, and density, or discuss data being clumped around the center. The end of this cycle then
forms the beginning of the second cycle pertaining to the development of reasoning of how to
use distributions for statistical inference. The unistructural, multistructural, and relational levels
coincide with the three levels of observing distributions asserted by Friel, O’Connor, and Mamer

A learner demonstrating distributional reasoning should reason with the relationships
among all five aspects of a distribution, so a critique that I have of this framework (Reading &
Reid, 2006) is that there seem to be many types of relational reasoning that a learner could
exhibit without demonstrating an ideal notion of distributional reasoning. For instance, coming to
conclusions through reasoning with center and spread without taking into account outliers may
lead to questionable conclusions. Reading and Reid's framework, along with the previous two
discussed (Bakker & Gravemeijer, 2004; Friel, O’Connor, & Mamer, 2006) give, in general
terms, an idea of understanding the relationship between the five distributional aspects, but do
not provide details illustrating what it means to have this coordinated view of distribution.
To bring together these frameworks for distributional reasoning in a hierarchical manner, I synthesized them into four levels of distributional reasoning: 1) Students exhibiting the lowest level distributional reasoning may reason with individual values in a data set, rather than any of the five aspects of a distribution; 2) Students at the next level draw conclusions about a distribution using a single aspect of a distribution, such as center without taking into account variation; 3) Students at the third level reason with multiple aspects, but do not relate their understanding of the aspects, such as noting one data set as being more spread out than another, but only taking into account a measure of center when drawing a conclusion; and 4) Students at the final level reason with the relationships between multiple aspects of a distribution. These four levels of distributional reasoning may help me to understand student reasoning as they develop informal inferential reasoning. Before students can make claims about a population of data, they first need to analyze the sample and sampling distributions. I next discuss research pertaining to how learners have used distributional reasoning to reason with distributions in graphical representations.

**Reasoning about graphical representations of distributions.** A graphical representation of a distribution may allow a learner to visualize aspects of a distribution such as skewness and density, but with this representation is the need for graph comprehension. Some research has examined the informal measures that learners use to reason about graphical representations of empirical distributions. Cobb (1999), for instance, investigated 29 seventh grade students and found that they applied informal measures of distributional aspects to analyze graphical representations of distributions. A common characteristic was the partitioning of the data into multiple sections that could be similarly described. The partitions were created when students found data values that were “bunched up” around a certain value. The use of the term
“bunched up” led to informal interpretations of the variation and shape of the distributions. Konold et al. (2002) found similar informal graph interpretations by seventh and ninth grade students describing “modal clumps”. The modal clumps were ranges of data values, typically centered around the mean, median, or mode, which contained a high proportion of the data points. Students used these modal clumps to give a description of the center of a distribution while also describing some aspect of variation.

Researchers have also investigated learners’ reasoning with formal measures of a graph. Cooper and Shore (2008) found that undergraduate students demonstrated difficulty determining aspects of a distribution such as mean, median, and variability when the distribution was represented as a histogram. Students also exhibited misconceptions with the distributional aspects of spread and density, often overvaluing the range of a distribution without taking into account other features of a distribution. The authors noted that it was troubling to see that 50% of the participants judged variability by focusing on the heights of the bars of the histograms. Cooper and Shore posited that the participants may have been confusing methods of determining variability in different graphical representations of data sets. The authors recommended that students may be better served by graphical representations that display all of the raw data, such as a stem and leaf plot. The histograms used in the study clustered much of the raw data together, which the authors asserted made it difficult for their participants to understand measures of center and variability. Once students are comfortable with those representations displaying all data points, the authors then suggest that students should be exposed to data sets with graphical representations of histograms with many shapes. This may support students’ understandings of the relationship between the center, variability, and shape of the distribution.

The progression from summarizing and describing the distribution of a single set of data
to working to comparing multiple sets of data has been asserted to provide the motivation and context for students to contrast and compare measures of center and variation (Konold & Higgins, 2003). The importance of an understanding of variation is amplified when comparing distributions, as there is not only variation within a distribution, but also variation between distributions. Comparing distributions may also be seen as a stepping stone to reasoning *with* a distribution and statistical inference (Watson & Moritz, 1999). Due to the increased sources of variation in addition to the demands of distributional reasoning and graph comprehension, many students struggle with the comparison of distributions even after extended instruction (Ben-Zvi, 2004).

Watson and Moritz (1999) examined the distributional reasoning displayed by Australian students in grades three through nine when comparing two distributions of data. The participants were given the distributions of the test scores of two classes and asked if the students in the classes scored equally well, or if one of the classes had scored better. This context was used to compare four pairs of data sets with varying distributions. The participants used both numerical and visual strategies, sometimes individually, sometimes in conjunction, in order to compare the sets of data. The study employed the SOLO framework, as discussed earlier in this review, and classified the knowledge displayed by the participants in two *unistructural*, *multistructural*, and *relational* (U-M-R) cycles. In the first cycle, *unistructural* responses tended to focus on the existence of “more” of one aspect of the graph, such as a student in one class having the highest score between classes. *Multistructural* approaches also looked into the idea of more points in the graph, but rather than compare highest scores, students found the total points earned by each class and compared these totals. This reasoning only took into account one aspect of the distribution, and would lead to differing conclusions depending on the particular distribution of
data. *Relational* responses were similar to the *multistructural*, in that they combined the use of comparing the total number of points in each class and the symmetrical shapes of the graphs to come to the conclusion that the classes scored equally.

Watson and Moritz (1999) suggested that proportional reasoning played a large role in participants' conclusions and constructed this second cycle to discuss proportional reasoning that students used when drawing their conclusions. *Unistructural* responses in the second cycle expressed some proportional reasoning, with responses discussing a class as having more higher scores for the amount in their class. *Multistructural* responses incorporated proportional reasoning while incorporating all data points by calculating the mean scores of each class. The final *relational* level incorporated the calculations of the mean scores in order to draw conclusions. The authors cited that a low proportion of the students drew valid conclusions from the use of the mean, and, as also suggested by Mokros and Russell (1995), past experiences calculating the value of the mean of a data set may interfere how the students apply this value in their reasoning. Watson and Moritz noted that both numerical and visual strategies of examining a distribution are critical and must be fostered by teachers, along with consideration of when one method may be more appropriate than another. The study demonstrated the importance of early experiences of informal reasoning in statistics and described the levels of reasoning that students may display when forming an understanding of this concept.

**Informal inferential reasoning via sampling activities.** The research pertaining to distributional reasoning that I have discussed has shown the ways that learners use distributions of data in order to make claims about those specific sets of data. A main tenet of statistical reasoning is making claims about larger unknown populations of data from a representative sample. The next body of research that I discuss uses similar empirical distributions of data to
not make claims about the specific data, but to make claims about the populations from which the data were drawn.

Sampling activities, such as those using repeated sampling and resampling methods, fall into the category of informal inferential reasoning. Repeated sampling refers to taking some number of samples from a population, while resampling uses only one sample from a population and some method of drawing resamples from that original sample.

Repeated sampling. Saldanha and Thompson (2002) conducted a teaching experiment with upper secondary students in which they took many samples from a variety of populations with known parameters using computer simulations. Goals of the experiment were to stress that “1) the random selection process can be repeated under similar conditions, and 2) judgments about sampling outcomes can be made on the basis of relative frequency patterns that emerge in collections of outcomes of similar samples” (p. 259). The authors stated that the activity helped some students to develop “a multi-tiered scheme of conceptual operations centered around the images of repeatedly sampling from a population, recording a statistic, and tracking the accumulation of statistics as they distribute themselves along a range of possibilities” (p. 261). A majority of the students did not focus on the distributions of sample statistics for inference and instead compared a single sample statistic with a population parameter. Saldanha and Thompson characterized these two conceptions of a sample in relation to the population and sampling distributions as multiplicative and additive, respectively. Those with an additive conception of the sample only viewed the part-whole relationship between the sample and the populations, with multiple samples representing multiple parts of this whole. The resemblance and relationship between the sample and population distributions was not a factor for those with this conception. Those with a multiplicative conception of the sample viewed the sample as a “quasi-
proportional, mini version” (p. 266) of the population. The sample can be used to approximate the distribution of the population, with an understanding that various samples' distributions may bear more or less resemblance to the distribution of the population. This multiplicative view may be critical to understanding the use of bootstrapping for informal inferential reasoning, since the proportion of each element in a sample is assumed to represent the proportion of those elements in the population.

Noll and Shaughnessy (2012) conducted a teaching episode with middle and high school students which involved repeated sampling with hands-on manipulatives. The study investigated the types of predictions that students made about a parameter of a population when reasoning with empirical sampling distributions. The authors created a conceptual lattice to characterize the types of reasoning that students exhibited when reasoning with sampling distributions. The lattice incorporated three hierarchical stages of reasoning: additive, proportional, and distributional. Additive reasoning was characterized by basing predictions on the frequencies of outcomes in the empirical sampling distribution. Predictions made with proportional reasoning involved proportional measures such as the mean. Distributional reasoning was exhibited by predictions made using multiple aspects of a distribution, such as center and variability. The authors found that students tended to predict narrow ranges for the population parameter, with a persistent tendency to rely on centers to make predictions. Noll and Shaughnessy asserted that to encourage classroom norms around interval estimation in statistics classes, teachers should create tasks that encourage students to provide an interval estimate that the students believe will capture the population parameter.

Pratt, Johnston-Wilder, Ainley, and Mason (2008) explored student thinking on the relationship between the sample and population with upper primary school students. Unlike the
study by Saldanha and Thompson (2002), the authors did not focus on the construction of an empirical sampling distribution. Instead, the study consisted of simulating the toss of a die with unknown outcomes and probabilities for the outcomes. The students accumulated the numbers of each roll, which placed the focus on the relationship between a growing sample and the population. The research was motivated by earlier research (Pratt & Ross, 2002) that indicated that students were proficient at describing short term randomness, identified as a local view, through observations such as the variations between subsequent trials, but had trouble expressing their understanding of the long term randomness, identified as a global view, such as the overall trend of a growing sample. Pratt et al. asserted that students may have trouble inferring this long-term randomness if the sample data is all that they have to work with, without the ability to draw more samples, or to add elements to samples. The study found that while the students wanted to generate more data in order to feel confident about their conclusions about the likelihood of the die's outcomes, they found the conclusions that students made from large amounts of data to appear to the students as no stronger than inferences made from smaller samples of data. The authors thought that this may be because even after rolling the die many times, the students still observed fluctuations, however small, in the relative frequencies of the outcomes of the die. The students may have been looking for the relative frequencies to at some point equal the expected proportions and from that point on, no longer change. This lack of convergence may have led the students to think that since the relative frequencies are always in flux, the size of the sample does not necessarily play a role in the strength of inferential claims.

These studies (Noll & Shaughnessy, 2012; Pratt, Johnston-Wilder, Ainley, & Mason, 2008; Saldanha & Thompson, 2002) indicate difficulties demonstrated by students when reasoning with samples and sampling distributions in order to make claims about the population.
Additive and multiplicative views of the relationship between the sample and population have bearing on my study and my participants’ understanding of the sampling distribution. The local and global views predicting randomness had an effect on the arguments used to make inferential claims. These views were discussed in the context of repeated sampling, but are just as relevant to resampling, which I will discuss next, as both activities incorporating either form of sampling aim for student understanding of the making of claims about the population. The tendency for students to focus on the center of a distribution rather than other aspects of the distribution influenced both the design and implementation of the instructional unit in my study.

**Resampling.** With the trend in statistics education towards data driven methods of teaching statistics, new curricula for introductory statistics courses have been created that emphasize the ideas of data creation, exploration and simulation (Garfield, delMas, & Zieffler, 2012; Pfannkuch, Forbes, Harraway, Budgett, & Wild, 2013; Tintle, VanderStoep, Holmes, Quisenberry, & Swanson, 2011). The CATALST curriculum (Change Agents for Teaching and Learning Statistics) was developed by Garfield, delMas, and Zieffler and explores the ideas of chance and models, simulation, and randomization via modeling activities in order to build students’ reasoning of statistical inference. The curriculum was taught as a teaching experiment in two tertiary courses and consisted of three units: chance models and simulations, models for comparing groups, and estimating models using data.

The first unit of the course covers material equivalent to what I have described earlier as repeated sampling. It is taught through the use of model eliciting activities, which are a focus of this study, and will be discussed later in the review. The second unit uses randomization methods, a form of resampling to compare groups of data. Randomization activities compare the difference in a statistic from two groups of data. The goal is to determine how likely it is to find
two groups that differ by this amount if the populations from which the groups were drawn are equal. The difference between statistics of two groups, such as the means, is first calculated. The data from each group is then pooled together and randomly assigned to one of the two groups. The same difference in a statistic from the two groups is calculated, and the process of pooling and random group assignment is repeated. The difference values are aggregated to construct the sampling distribution of the difference between the statistics of the two groups of data when the populations of data have the same parameter. From one sample, many resamples were constructed and used to construct a sampling distribution. From there, informal inferences can be made regarding the likelihood of the original sample. The third unit of the CATALST curriculum (Garfield, delMas, & Zieffler, 2012) introduces the bootstrap method as a tool for estimating population parameters.

Researchers analyzing the preliminary data of two teaching experiments with the CATALST curriculum noted that even in an introductory statistics course, students are capable of developing an understanding of statistical inference. The material covered in this course was much different than the traditional introductory statistics courses based on numerical approximations and the normal distribution. The study found that the students performed as well as those enrolled in a traditional course. To determine this, the study used the MOST (Models of Statistical Thinking) and GOALS (Goals and Outcomes for Assessment and Learning of Statistics) instruments, the latter of which was used in my study. The coursework used TinkerPlots (Konold & Miller, 2014) software, which I will discuss later in this literature review, and determined that students could be taught to “really cook” (Garfield, delMas, & Zieffler 2012, p. 884) and create representations of statistical inference. A similar curriculum consisting of randomization and bootstrapping content was implemented in high schools and introductory
statistics course at the university level throughout New Zealand (Pfannkuch, Forbes, Harraway, Budgett, & Wild, 2013). In line with the CATALST curriculum, the authors found that bootstrapping and randomization methods using software capable of dynamic visualizations have the potential to transform the learning of inference.

Much of the content in these curricula is similar to my study, as the curricula covered repeated sampling, resampling, and bootstrapping methods. While ideas of repeated sampling and the resampling method of randomization for comparing groups in the CATALST curriculum are elicited in model eliciting activities, the same is not true for the third unit covering estimation for the bootstrapping method. The curriculum materials indicate that the students were told of the bootstrapping method, and did not have the opportunity to construct the method on their own. The curriculum developed in New Zealand by Pfannkuch et. al (2013) follows the similar path of teaching the concept of bootstrapping without allowing the students to construct the concept on their own. I suggest that this lack of eliciting the concept of bootstrapping is a gap in the research literature. My study continued in the direction of the CATALST curriculum, while exploring how students develop understanding of resampling and in particular, the bootstrap method.

**Technology Use in Statistics Education**

Traditional approaches to teaching statistics have often overly focused on mathematical procedures rather than statistical thinking or reasoning (Ben-Zvi & Garfield, 2008). Advances in technology such as graphing calculators, web applets, and a number of software programs can provide opportunities to move beyond this procedural focus on statistics and to an understanding of core concepts of statistical thinking. Technology can support student learning through the automation of calculations, emphasis on data exploration, visualization of abstract concepts, and use of simulations as a pedagogical tool (Garfield & Ben-Zvi, 2008). With the use of technology,
Chance, Ben-Zvi, Garfield, and Medina (2007) suggested that resampling methods, which may be more intuitive to students, can compete with the commonly taught inferential procedures largely used in statistics classrooms. Technology in some form is in use in the majority of introductory statistics classrooms (Garfield, Hogg, Schau, & Whittinghill, 2002).

The statistics education literature has noted the difference between technologies that are for *doing* statistics and those that are for *teaching* statistics. Those that are for *doing* statistics are often considered black boxes, where the user inputs values related to their problem and the technology calculates an answer. Examples of black box technologies are standard graphing calculator programs and *Minitab* software (Minitab 17 Statistical Software, 2010). In each case, if the user wants to perform a specific hypothesis test, the user inputs data values and with a few clicks performs a hypothesis test and is given a p-value. Researchers have suggested that there are limitations in this approach of technology use in the learning of statistics (Meletiou-Mavrotheris, 2003). Moore cautioned that in statistics education, we are “teaching our subject and not the tool” (1997, p. 135). Black box software used by statisticians to *do* statistics may not promote student learning of statistics, only how its calculations can be executed. Olive and Makar (2010) asserted that technology should allow the teaching of statistics to be visual, dynamic, and interactive with students engaging in experimentation with data rather than computations. Software programs such as *Fathom* (Finzer, 2001) and *Tinkerplots* (Konold & Miller, 2014) are in line with Olive and Makar’s recommendations, and have been used in recent studies (delMas, Garfield, & Zieffler, 2014; Makar & Confrey 2004).

Makar and Confrey (2004) viewed *Fathom* (Finzer, 2001) as supporting their participants' statistical learning by allowing the participants to fluidly investigate the data in various graphical representations because of the ease of producing graphs within the software, and the ability to
create transparent simulations to test conjectures. Their study worked with in-service secondary teachers in the context of professional development workshops and a summer institute and investigated data regarding their state's high stakes testing. The authors used *Fathom* to aid in the participants’ struggles with the ideas of variation within a distribution, variation between two distributions, and interpreting the difference between these two forms of variation.

Watson and Chance (2012) asserted that the *TinkerPlots* (Konold & Miller, 2014) software, which similarly to *Fathom* (Finzer, 2001) was created for the learning of statistics, provides an easy to use means for constructing investigations, displaying graphical representations of data, exploring concepts in resampling and randomization procedures. The CATALST curriculum (Garfield, delMas, & Zieffler, 2012) incorporated *TinkerPlots*. To understand its effectiveness in the curriculum, select students participated in problem based interviews after the first five weeks of the CATALST curriculum. The analysis indicated that *TinkerPlots* offered a memorable visual medium that supported their development of statistical reasoning (delMas, Garfield, & Zieffler, 2014). Similar acclaim for the software was offered by Watson and Donne (2009) who described the affordances of *TinkerPlots* for exploring students' understanding as the flexibility of representation, the speed of analysis, and the exposure of varying levels of student understanding.

These evaluations of the software describe the many benefits of visual, dynamic and interactive software with the purpose of teaching statistics rather than only doing statistics. Since this study involved the drawing of many samples and comparing of changing distributions, this form of software seems key to aiding in student understanding. Garfield, delMas, and Zieffler (2012) viewed *Fathom* (Finzer, 2001) and *TinkerPlots* (Konold & Miller, 2014) as suitable tools for teaching statistics with data simulations. Although *TinkerPlots* was developed for use in
elementary and middle schools, Garfield, delMas, and Zieffler chose to use *TinkerPlots* at the university level. *TinkerPlots* had features not seen in *Fathom*, such as samplers and spinners to easily visualize data collection. With this in mind, I chose *TinkerPlots* for this study. I investigated how these visualizations of data collection aided student understanding of sampling as students move from using hands-on manipulative to computer simulations.

**Modeling**

The main focus of analysis for this study is the models of resampling and inference that the students create while engaging in two model development sequences. Models are:

- conceptual systems (consisting of elements, relations, operations, and rules governing interactions) that are expressed using external notation systems, and that are used to construct, describe, or explain the behaviors of other system(s) – perhaps so that the other systems can be manipulated or predicted intelligently. (Lesh & Doerr, 2003, p. 10)

Models are systems that can be used to describe another system for a specific purpose (Lesh & Fennewald, 2010). Teaching and learning from a modeling approach shifts the focus of an activity from finding an answer to one particular problem to constructing a system of relationships that is generalized and can be extended to other situations (Doerr & English, 2003). Students' mathematical models are useful for research since they provide a means for investigating students' developing knowledge. Modeling activities provide researchers with access to the ways of student thinking that produce models, such as the intermediate steps and processes that contributed to the final result (Lesh et al., 2000).

Model development sequences focus on higher order understandings of a small number of big ideas, rather than a large number of smaller facts and skills (Ärlebäck, Doerr, & O'Neil, 2013; Lesh, Cramer, Doerr, Post, & Zawojewski, 2003). These sequences consist mainly of three
forms of activities related to models: model eliciting activities (MEAs); model exploration activities (MXAs); and model application activities (MAAs). Throughout the sequence students have the opportunity to present, discuss, and reflect on their experiences in each form of activity.

**Model eliciting activities.** Model development sequences generally begin with a model eliciting activity that encourages students to generate descriptions, explanations, and constructions in order to reveal how they are interpreting and mathematizing problematic situations. The activities are designed to emphasize student-to-student interactions, with participants generally working in groups of three to four students (Lesh et al., 2003). The most important criterion of these activities is that they explicitly reveal the students' development and construction of their models. This will allow for the investigation of the students' developing mathematical constructs (Lesh et al., 2000). In order to meet this criterion, Lesh, et al. (2000) recommended the following six principles to be used when designing model eliciting activities: 1) the reality principle; 2) the model construction principle; 3) the self-evaluation principle; 4) the model externalization principle; 5) the simple prototype principle; and 6) the model generalization principle.

The *reality principle* states that students should be able to make sense of the situation in the activity and be able to draw on both their school mathematics abilities and their real life sense making abilities (Lesh et al., 2000). The task should allow the students to investigate these real life sense making abilities (which may differ from the teacher's) and not be forced to pursue the teachers' notions of the correct way to proceed in the activity. The second principle, the *model construction principle*, states that students must recognize the need to construct a model to complete the activity, rather than produce only an answer to a question asked by someone else. The task should focus on patterns and irregularities and force the students to construct, describe,
explain, manipulate, or predict elements of a structurally significant system (Lesh et al., 2003).

The self-evaluation principle states that the task should allow students to assess when their responses need to be improved, refined or extended. The purpose of the task should be clear so that the students can judge their responses as good enough for the situation, without guidance from the teacher (Lesh et al., 2000; Lesh et al., 2003). The fourth principle, the model externalization principle, also called the documentation principle, states that the tasks must not be only thought eliciting, but thought revealing. Lesh et al. (2000) suggested that the task should have the students create an “audit trail” that not only lists the task's final solution, but also the goals, possible solution paths, and objects, relations, operations, patterns, and irregularities that the students took into consideration while working towards their solution. The simple prototype principle asks if the situation of the task is as simple as possible, while allowing the need to construct a significant model. A goal of the model eliciting activity is to create a significant, but compact model that can then be used to interpret other structurally similar situations (Lesh et al., 2003). This ties into the final principle for designing model eliciting tasks, the model generalization principle, which states that the task should encourage the creation of models that not only apply to the single specific situation of the task; the models should be able to be extended to use in a broader range of situations (Lesh et al., 2003). Next in a model development sequence, students may take part in activities that explore the inner working of their model, or apply the models to new situations.

Model exploration activities. Once students have formed their models, exploration activities focus on the mathematical structure of their models, along with the pros and cons of various representations of the models (Ärlebäck, Doerr, & O'Neil, 2013; Doerr, Arleback, & O'Neil, 2013). These activities often do not rely on concrete materials and use computer
graphics, diagrams, or animations in order to develop a powerful representation system for the students to make sense out of the situation (Lesh et al., 2003). In model eliciting activities, students are encouraged to invent their own language, diagrams, metaphors, and notation systems in order to express their thinking on construction of their model. During model exploration activities, since the goal is to introduce students to conventions that have taken mathematicians and statisticians many years to develop, it is generally not expected for students to come up with this information on their own (Lesh et al., 2003). This more in-depth understanding of their model may allow the students to complete more complex versions of the situations originally seen in model eliciting activities, and found in model application activities.

**Model application activities.** The models, or tools, created in model eliciting activities are then used to deal with problems in model application activities that may have been too difficult for the students to complete prior to the model development sequence (Lesh et al., 2003). The goal for the models constructed in the model eliciting tasks are for them to be easily generalizable to new situations, but significant changes to the original model are generally needed to complete model application activities. In contrast to model eliciting activities, model application activities are often completed by students alone. This can lead to the view of model application activities as a post-test, with the model eliciting activity as a pre-test for a big idea in a unit (Lesh et al., 2003).

**Summary**

I began this review by discussing informal inferential reasoning and distributional reasoning. This discussion focused on the reasoning of students when making inferential claims from data and influenced my analysis of evidence of student reasoning. Since this study also involved the use of statistical software, I discussed forms of software for doing statistics and
those for learning statistics. Software designed to promote the learning of statistics was used in this study to aid students in their investigations of reasoning with data. Lastly, model development sequences have influenced the design of this study. Modeling allows students to make sense of a situation and develop not only correct answers, but a system to determine that answer. Since I studied the development of understandings of statistical inference through a bootstrapping approach, model development sequences allowed me to observe the sense that the students made of the various situations in activities, and the development of reasoning evidenced by the systems that they developed to investigate the tasks.
Chapter Three - Design and Methodology

In this study, I investigated students' learning of bootstrapping methods by examining: 1) Reasoning that developed as students moved from repeated sampling methods to bootstrapping methods; 2) How students developed reasoning about bootstrapping methods in order to make informal inferences about a population of data; and 3) Student reasoning that was revealed and supported by moving from using hands-on manipulatives to computer simulations during repeated sampling and bootstrapping activities.

Research Design

This study was a qualitative case study, with an intervention, investigating the development of student reasoning of sampling and informal inference. The intervention was a one-month, eight-class-session teaching experiment, enacted in four introductory statistics classes at the high school and community college levels, consisting of an instructional unit developed as two model development sequences. I designed the first model development sequence to elicit and develop participants’ models of repeated sampling and inference. The second model development sequence adapted participants’ models from the first sequence in order to elicit and develop the concepts of bootstrapping and inference. During both model development sequences, hands-on manipulative and TinkerPlots software were used to simulate data. I chose TinkerPlots because of the software’s dynamic representations for the processes of repeated sampling and resampling and the speed of the sampling process in comparison to hands-on manipulatives.

During the teaching experiment, I collected written classwork from all participants and videotaped full class discussions and presentations. I followed four focus groups of students (one from each class) and videotaped their group discussions in order to document the groups’ model
development. From each of these focus groups one student participated in three 30-minute interviews. These interviews helped me to gain insight into the groups’ model development by questioning the participants on aspects of their developing models that had not been captured by other data sources.

**Settings and Participants**

I recruited two instructors, each teaching two sections of an introductory statistics course. From the four classes, I recruited 68 students (all students participated) to participate in the study. The first instructor, Brenda, was an instructor at a community college. I choose her as a participant due to her 12 years of experience teaching statistics, interest in moving her courses away from traditional statistics instruction, and her willingness to devote four weeks of class time to the study. I recruited all students in her two sections (n = 18, 15) as participants in the study. The classes met twice each week, for 75 minutes each class session.

The second instructor, Allen, was a high school teacher with one year of experience teaching statistics, but many years of mathematics teaching experience. He was similarly chosen as a participant due to interest in nontraditional statistics instruction. The course that he taught was for upper secondary school students, but not an advanced placement course tied to college credits. This allowed him more leeway in the material that he covered and he agreed to devote four weeks to the study. I recruited all students in his sections (n = 22, 13) as participants in the study. The classes met every other weekday for 80 minutes each class session.

While the participants in the study are from a high school and community college, they are all students in introductory statistics courses. Since it was not likely for the students to have had coursework experience in formal statistical inference, I made the assumption that the students had a similar understanding of inferential statistics. I tested this assumption by
quantitatively analyzing and comparing pre-test results of the participants in the two settings.

**Model Development Sequences**

**Overview of sequences.** Groups of three to four students participated in an instructional unit over eight class sessions that was constructed as two model development sequences (Figure 4). The first model development sequence elicited and developed participants’ models of repeated sampling and inference. In this model development sequence, populations of data were available for the groups of students to collect many samples of data. The first model development sequence began with a model eliciting activity in order to elicit students’ initial ideas about repeated sampling and making informal inferences from empirical sampling distributions. Model exploration activities investigated how changes in aspects such as the number of collected samples affected groups’ models of sampling and inference. Model application activities investigated how groups of students could adapt their models of sampling and inference to new situations, such as the comparison of multiple populations.

The goal of the second model development sequence was for students to adapt their models of sampling and inference to construct the resampling process of bootstrapping. In this model development sequence, populations of data were not available for the groups of students to collect many samples of data. The second model development sequence began with an activity requiring the students to make inferences about a population of data and to adapt their previous models to work with only one sample of data. Participants did not have the option of repeatedly sampling from the population, which necessitated the construction of models of resampling from the one given sample, such as the bootstrapping method. A model exploration activity investigated how changes in aspects such as the contents of the one original sample affected groups’ models of resampling and inference. A model application activity investigated
how groups of students could adapt their models of resampling and inference to compare multiple populations of data.

Figure 4. Overview of the activities in the two model development sequences.

First model development sequence. The first model development sequence intended for students to develop models to make inferential claims from empirical sampling distributions constructed by repeatedly sampling from an available population. Table 1 indicates the models that I intended to elicit, explore, and apply in each activity in this model development sequence.
### Table 1

**Intended Models in the First Model Development Sequence**

<table>
<thead>
<tr>
<th>Activity</th>
<th>Summary of Activity</th>
<th>Intended Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 MEA</td>
<td>Students used a cup of coins to simulate the possible outcomes when an octopus predicts the winners of eight basketball games.</td>
<td>Elicits a model that students use to generate data and use the relative frequency of the outcomes in the data to determine which outcomes are most likely to occur.</td>
</tr>
<tr>
<td>2 MXA</td>
<td>Students used <em>TinkerPlots</em> to simulate the possible outcomes when an octopus predicts the winners of eight basketball games.</td>
<td>Explores the dotplot representation of increasing numbers of samples of simulated data to determine which outcomes are most likely to occur.</td>
</tr>
<tr>
<td>3 MAA</td>
<td>Students used a bin filled with thousands of multicolored beads, representing brands of sneaker purchases by university students, to simulate the likely outcomes when randomly sampling groups of 20 students.</td>
<td>Applies their model to support or oppose claims regarding the population from which the data was drawn.</td>
</tr>
<tr>
<td>4 MXA</td>
<td>Students used <em>TinkerPlots</em> to simulate the likely outcomes for brands of sneaker purchases when randomly sampling groups of 20 students.</td>
<td>Uses dotplot representation of increasing numbers of samples of simulated data to explore the relative frequency of outcomes and the proportion of outcomes considered to be likely.</td>
</tr>
<tr>
<td>5 MAA</td>
<td>Students again used a bin filled with thousands of multicolored beads to simulate the likely difference in sneaker purchases between two brands of sneakers when randomly sampling groups of 20 students.</td>
<td>Applies their model to compare two populations and use simulated data to support or oppose claims regarding the populations from which the data was drawn.</td>
</tr>
</tbody>
</table>

The modeling-eliciting activity (MEA) told a brief story of an octopus that predicted eight of eight soccer matches correctly during the 2010 World Cup (Appendix C). Local zookeepers planned to have their octopus predict the winners of eight upcoming basketball games. Groups of students were asked to write a letter to the newspaper with the range of
outcomes that they thought were likely for the octopus to correctly predict, including a description of the methods that they used to come to this conclusion. The students were given a cup of coins to elicit the idea that the octopus was guessing the outcome of each basketball game, which could be simulated by flipping a coin. Through flipping eight coins many times, students could observe the patterns of outcomes, with the outcomes occurring most often being considered the outcomes most likely to occur.

A homework assignment then discussed Paul the octopus, who predicted eight of eight soccer matches correctly during the 2010 World Cup. The students were told that a blog claimed that Paul must be psychic and asked them to comment on the blog post. In the MEA, students predicted likely future outcomes, but in this task, students now evaluated what it meant for an unlikely outcome to occur (Appendix C).

The next activity was a model exploration activity (MXA), which investigated the same context with Ophelia as the MEA, but students now used TinkerPlots (Konold & Miller, 2014) to simulate outcomes rather than the coins (Appendix C). I created an environment in TinkerPlots that contained a sampler, results, and dotplot window (Figure 5). The sampler window in TinkerPlots (shown in the upper left) used a spinner with half of the area marked right, the other half wrong. This corresponded to the flipping of one coin to predict the octopus’ answers. The results window (shown in the lower left) showed the outcomes generated by the spinner in sets of eight predictions. The dotplot window (shown on the right) displayed the number of right predictions in each set of eight. Each student ran the simulation on an individual computer. The activity asked the students to compare their individual dotplots and inferences with their group members’. The activity guided the students to explore how the collection of an increasing number of samples affected the conclusions that they drew from the data. With the dotplot
representation of the data, students had the opportunity to visually explore the shapes of the distributions of data collected by fellow group members, and the changing shape of their own distribution of data as an increasing number of samples is collected. Figure 5 shows a collection of 50 samples; however, when the students opened the TinkerPlots environment, they began with no collected samples.

![Image of TinkerPlots environment](image)

**Figure 5. TinkerPlots environment for the Ophelia MXA.**

A homework assignment then returned to the question from the MEA and asked individual students to again write a letter to the newspaper with the range of outcomes that they thought were likely for the local octopus Ophelia to correctly predict and include a description of the methods that they used to come to this conclusion. I expected that the use of TinkerPlots would have an impact on values of this predicted range and how the range was determined. I designed this homework assignment to inform my third research question regarding the students’ reasoning that was revealed and supported by moving from hands-on manipulatives to technology.
The third activity was a model application activity (MAA) in which the students adapted their models of sampling and inference. In the MEA, students generated a sample by assuming the probability of an octopus choosing the winner of a game was 1/2. The coins did not represent the population of predictions, but could be used to generate each sample predictions. In this MAA, the students did not know the actual probability of selecting each element of their sample. There was now a manipulative to represent each element of the population. The students were told that in 2013, 35% of all sneakers sold globally were Nikes. A marketing director from Nike contacted the students to help him determine if this global trend of Nike sneaker sales held true for the students at a local university. The marketing director wanted to discuss sneaker sales for the local university’s students at an upcoming meeting and wanted to know if it was reasonable to claim that about 7 in 20 of the local university’s student sneaker purchases were Nikes. The groups of students were given a bin containing thousands of multicolored beads. They were told that each bead in the bin represented a sneaker purchase by a student from the university and every clear bead was a Nike sneaker purchase. With the bin of beads, the students were given a spatula with 20 holes just smaller than the size of the beads (Figure 6). This enabled the students to collect samples of 20 beads quickly if they chose to do so.
The fourth activity was an MXA that continued in the context of Nike sneaker purchases using TinkerPlots (Konold & Miller, 2014) to allow the students to more quickly take many samples and explore the proportion of all outcomes that were in their ranges of likely outcomes. Just as in the previous activity using TinkerPlots, I created the environment with a sampler, results, and dotplot window. The sampler window contained the same number of balls as there were beads in the bin. The balls were labelled with the letters N, A, and O, where N represented Nike sneaker purchases, A represented Adidas, which were used in a later activity, and O represented other brands of sneaker purchases. The sampler window was setup to collect 20 balls at a time. Figure 7 shows the results of collecting 500 samples of 20 sneaker purchases; however, as in the previous TinkerPlots activity, the TinkerPlots environment contained no collected samples when first opened by the students. Each student had their own TinkerPlots environment in order to compare simulated data and conclusions with their group members. The MXA guided the students to explore the shape of the distribution for the number of Nike sneaker purchases in

Figure 6. Bin of multi-colored beads for the Nike sneakers MXA and MAAs.
their samples of 20 sneaker purchases, how these shapes varied between their group members’ dotplots, and how the shape of the distributions varied as more samples were collected.

![Figure 7. TinkerPlots environment for the Nike sneakers MXA.](image)

The final activity of the first model development sequence was an MAA. Groups of students were to adapt their models of sampling and inference to compare two characteristics in a population. In the previous activities, groups of students drew inferences from one sampling distribution. In this MAA, groups were to adapt their models to make inferences by comparing two empirical sampling distributions. The students were told that Adidas’ global share of the sneaker market is 20%. Nike is interested to know if this difference of about 15% between Nike and Adidas sneaker purchases was true for sneaker purchases of local university students. The students were told that in addition to the clear beads in the bin representing Nike sneaker purchases, the blue beads were Adidas sneaker purchases. Nike asked the groups of students to determine the likely range for the difference in percentage between Nike and Adidas sneaker purchases for local university students.
Second model development sequence. The key aspect of the first model development sequence was that groups of students had access to an available population. Many samples could be collected from this population in order to construct an empirical sampling distribution and draw inferential claims. The major difference between the two model development sequences was that in the second there was no longer an available population. Groups could not collect many samples, which necessitated the groups to develop a means to construct an empirical sampling distribution from resampling from one original sample of data. The groups of students had only one sample that had already been drawn from the population to use to make inferential claims (see Figure 3). This is a more realistic scenario for inference than that used in the first model development sequence. It is often impractical to draw many samples from a population and methods of inference from one sample are needed. Table 2 indicates the models that I intended to elicit, explore, and apply in each activity in this model development sequence.
Table 2

**Intended Models in the Second Model Development Sequence**

<table>
<thead>
<tr>
<th>Activity</th>
<th>Summary of Activity</th>
<th>Intended Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 MEA</td>
<td>Students then used only one sample of mixed nuts, represented by marked craft sticks, in order to make claims about the proportion of peanuts in the entire brand of mixed nuts.</td>
<td>Elicits a model that is used to resample from a sample, with replacement, in order to construct a bootstrap sample of the same size as the original sample and use the distribution of these bootstrap samples to determine which are most likely to occur.</td>
</tr>
<tr>
<td>7 MXA</td>
<td>Students used <em>TinkerPlots</em> to resample from their sample of mixed nuts to simulate the proportion of peanuts in a brand of mixed nuts.</td>
<td>Explores dotplot representations of simulated data to determine likely outcomes and compares of reasons with simulated data generated from different original samples.</td>
</tr>
<tr>
<td>8 MAA</td>
<td>Students compared two samples of mixed nuts from different brands and used resampling in order to make claims about which brand contained the lower proportion of peanuts.</td>
<td>Applies the model to compare two populations and make claims about the populations from which the data was drawn</td>
</tr>
</tbody>
</table>

The first activity of the second model development sequence was the Mixed Nuts MEA (Appendix C). Groups of students were asked to help a manager in a grocery store predict the percentage of peanuts in a certain brand of mixed nuts. The students were given a sample of mixed nuts in the form of seven craft sticks marked with a ‘P’ for peanut and 18 not marked to represent other kinds of nuts (Figure 8).
The manager planned to buy a large shipment of mixed nuts, but thought that her customers preferred fewer peanuts. From this one sample of mixed nuts, the manager asked the students to determine a likely range for the percentage of peanuts in the entire brand of mixed nuts. This activity was in contrast to the MEA in the first model development sequence, since the students did not have the option to take many samples of mixed nuts to draw their conclusions.

The next activity was an MXA (see Figure 4) that continued to investigate in the context of mixed nuts, from the initial MEA in the sequence, with TinkerPlots (Konold & Miller, 2014). The TinkerPlots environment was set up to resample from the original sample of 25 nuts and collect bootstrap samples of 25 nuts. This allowed the students to quickly collect many samples, which was time consuming with the sticks. Just as with all previous MXAs, each student had their own TinkerPlots environment in order to allow group members to compare their simulated data and inferential claims.
The final activity was an MAA (Appendix C). The purpose of the activity was for students to adapt their models of bootstrapping and inference from one population to the comparison of two populations and two empirical bootstrap sampling distributions. The store manager decided to order a second sample of mixed nuts, but this time from a different brand. She wanted to know which brand had the smaller percentage of peanuts. From those two samples of mixed nuts, she wanted the students to find the likely range of the differences in the percentages of peanuts for the two brands. The students were given two bags of sticks, marked A and B, representing samples of mixed nuts from two brands. The sticks marked with a ‘P’ again represented peanuts and the unmarked sticks represented other types of nuts. Students had access to the TinkerPlots (Konold & Miller, 2014) setup from the previous activity, and were able to modify the number of peanuts in the TinkerPlots sampler to coincide with their two samples.

**Data Collection**

**Before the model development sequences.** Before enacting the model development sequences, I met with the instructors individually to collaboratively plan the implementation of
the activities. We also discussed the selection of participants for the focus groups, with the instructors choosing participants who would be vocal regarding their thinking processes and likely to attend every class session.

All participants (n=68) took a written pre-test prior to the unit to measure their understandings of informal inference (Appendix A). The function of the pre-test instrument was to test my claim of homogeneity of the high school and community college participants’ backgrounds in statistics and inferential reasoning. The instrument was a subset of the Goals and Outcomes Associated with Learning Statistics (GOALS) instrument that was designed to develop and evaluate innovative lessons for teaching introductory statistics. I chose a subset of 13 multiple choice items of the instrument to contain items relevant to the material discussed in the instructional unit and informal inferential reasoning that may have had an effect on participants’ model development. The items covered content such as the relationship between the sample and its population, the comparison of samples to make claims about their populations, empirical sampling distributions, variability, and resampling.

**During the model development sequences.** During each of the four class sessions, one focus group was videotaped during all group work. I videotaped the front of the classroom during whole class discussion and group presentations to capture group presentations, group work written on the front white/chalk board, and the instructors’ interactions with the groups as they presented material to the class. I videotaped these whole class discussions to gain insight into participants’ developing models of sampling and inference that I may not have captured in the focus groups.

I collected written work from all participants. This written work consisted of scrap work written on the distributed activity sheets, homework assignments, and documents produced
during the activities, such as the letters that groups wrote to the newspaper. Just as with the videotaping of whole class discussions, collecting this written work from all participants allowed me to analyze developing models of sampling and inference that were not demonstrated in the focus groups.

To gain insight into the model development of students in the focus groups beyond what was captured by the videos and written work, I recruited one participant from each focus group to participate in three 30-minute interviews throughout the instructional unit. The instructors suggested the participants to choose based on the participants being the most vocal about their thoughts during class. The interviews occurred at the end of the first model development sequence, after the first activity of the second model development sequence, and after the two model development sequences. I videotaped each of the interviews and collected any written documents created during the interviews. The interviews gave me the opportunity to delve deeper into the thinking used by the students while completing the activities in order to understand aspects of the development of their models not conveyed during class. In Table 3, I list the participants in each focus group (pseudonyms), their instructor’s name, and the students who participated in the interviews.
Table 3

*Focus Group Participants*

<table>
<thead>
<tr>
<th>Allen’s Classes</th>
<th>Brenda’s Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Focus Group A:</strong> Aaron</td>
<td><strong>Focus Group C:</strong> Karl</td>
</tr>
<tr>
<td>Beth</td>
<td>Lena</td>
</tr>
<tr>
<td>Cara*</td>
<td>Megan*</td>
</tr>
<tr>
<td></td>
<td>Nate</td>
</tr>
<tr>
<td><strong>Focus Group B:</strong> Frank*</td>
<td><strong>Focus Group D:</strong> Randy*</td>
</tr>
<tr>
<td>George</td>
<td>Susan</td>
</tr>
<tr>
<td>Henry</td>
<td>Ted</td>
</tr>
</tbody>
</table>

*Participated in individual interviews*

The instructors were also debriefed after each class session to discuss how the class and activities went that day. I audiotaped each interview. An interview protocol for these debriefings is included in Appendix B. The purpose of the debriefing was to discuss possible student thinking that may not have been captured by the videos or collected written work, discuss how the lesson worked in the classroom, and possible ways of improving the lesson and later lessons in the instructional unit. After each lesson, I wrote a memo with my thoughts and observations of the lesson.

**After the model development sequence.** The students took a post-test (Appendix A) identical to the pre-test in order to measure changes in students’ informal inferential reasoning. Items on the pre/post-test directly related to the content of the instructional unit, such as the representativeness of the sample, were the focus of the quantitative analysis.

**Data Analysis**

**Pre/post-test analysis.** I compared the pre-test results of the participants in the two settings with one-way ANOVA to determine if there were differences between the mean scores on the instrument from participants in the high school and community college setting. Each item
of the pre-test was scored 1 for correct, 0 for incorrect, and summed for each participant. I did this comparison to gain evidence of difference in the backgrounds of participants’ statistical reasoning between the high school and college settings. If there was a difference, model development for each setting may need to be analyzed separately to take into account these differences. The pre-test and post-test results were analyzed with paired t-tests to determine if there were statistically significant increases in the mean scores for all participants, participants in each setting, and each of the four classes separately. Only data from students who had completed both the pre-test and the post-test was used in this analysis.

**Qualitative analysis.** I analyzed the videos of the four focus groups, whole class discussions, and interviews from each class so that I could reconstruct the development of the reasoning that were created by the participants. This allowed me to draw conclusions regarding how their reasoning of repeated sampling and bootstrapping developed. To do this, I grouped the data collected into three levels: whole class data, group data, and individual data.

**Whole-class data.** The whole-class data consisted of audiotaped discussions of the implementation of the instructional unit with myself and the two instructors, the videotapes of group presentations and whole-class discussions during the activities, audiotaped debriefing of the instructors after class, written student work collected during class, written student homework, and memos containing my observations of the class and student thinking. The primary data sources were produced by the participants and involved student thinking. The data from the instructors were a secondary source. I analyzed the in-class videos and written work for evidence of student thinking, common errors, quantities and their relationships, and the representations used throughout the activities. I used open coding on these data and searched for emerging themes. With this information I wrote a summary of the statistical thinking displayed by the class
as a whole (including both focus group and non-focus group participants) to gain perspective on
the statistical thinking of introductory statistics students. I used the secondary data sources from
the instructors to solicit their observations of student thinking in the classrooms and thoughts
about the learning that they intended to occur during the lessons versus what they thought had
occurred. I added aspects of these conversations to the summaries of the classes’ statistical
thinking when relevant.

*Group data.* I videotaped the four focus groups working through each activity and I
collected their written work. With these data, I conducted a four-phase analysis. First, for each
focus group I examined the videos of the focus group and their written work for evidence of
informal inferential reasoning, quantities used and the relationships between these quantities, and
the representations and explanations used in their arguments to construct summaries of their
thinking on a per activity basis. I used these summaries to construct a trajectory for how each
focus group’s thinking developed across the instructional unit. Second, I searched across the
summaries and trajectories of the four focus groups to identify emerging themes on the
development of sampling and informal inferential reasoning. Third, I examined how the whole
class and individual data sources informed the focus group summaries. The basis for this study is
the development of thinking in the focus groups, with other data sources collected to gain a
greater understanding of this development. Fourth, I compared the model development across the
four focus groups to understand the differences and similarities between each group’s
demonstrated reasoning of sampling approaches and informal inferential reasoning.

*Individual data.* For each focus group, I videotaped interviews with one of the group
members either two or three times (based on participant availability) throughout the instructional
unit and gathered written work produced during this interview. These data were used while
analyzing the focus-group data in order to fill in student thinking that occurred during the activities, but may not have been captured by other sources. The first interview took place towards the end of the first model development sequence. I chose this timing to investigate interviewees’ understanding of repeated sampling and the sampling distribution that I may have missed in class. The second interview occurred during the second model development sequence after the participants had completed the model eliciting activity involving resampling and the proportion of peanuts in a certain brand of mixed nuts. This activity was the participants’ first exposure to resampling in the model development sequence. Interviewing focus group members at this time allowed me to investigate their initial reasoning of resampling that may not have been evident in the videos and written classwork. The third interview took place at the end of the model development sequences and allowed me to further investigate the participants' reasoning of bootstrapping with one or two populations, the relationship between the inferences made with resampling and repeated sampling, and other pertinent topics that arose throughout the instructional unit.
Chapter Four – Results

This study was a qualitative case study, with an intervention, investigating the development of student reasoning of sampling and informal inference. The intervention was a teaching experiment that was enacted in four introductory statistics classrooms at the high school and community college levels. The teaching experiment consisted of two model development sequences that elicited and developed students’ models of sampling and the resampling process of bootstrapping. Hands-on manipulatives and data simulations with TinkerPlots (Konold & Miller, 2014) were used to construct empirical sampling distributions and investigate the inferences that could be drawn from the data. The first model development sequence elicited and developed models of sampling and inference. Groups of students made inferential claims from empirical sampling distributions that were constructed by repeatedly sampling from a population. For the second model development sequence, the population was no longer available from which to repeatedly sample. Groups of students had only one sample from the population and some groups constructed the resampling method of bootstrapping to draw inferential claims.

Groups of students followed a multiplicity of paths in the development of their models of sampling and inference. The focus of my analysis was on the model development of four focus groups of students, with one group coming from each of the four classes. Groups of student constructed models of sampling and inference that they used to collect samples, aggregate the samples to form an empirical sampling distribution, and use the aspects of this distribution of samples to make claims about the population from which the samples were drawn. I discuss models of sampling and inference constructed by groups outside of the focus groups if the models contrast with those of the four focus groups. I summarize and categorize groups of students’ models for each activity and trace the development of the focus groups’ models.
throughout the unit. I discuss how each of these main categories of models of sampling and inference progressed through the instructional unit, sometimes intersecting, and sometimes splitting into new models of sampling and inference.

The trajectories of the groups’ models of repeated sampling and bootstrapping allow me to address my three research questions. The first research question addresses what student reasoning developed as they moved from repeated sampling methods to bootstrapping methods. This is addressed while summarizing the development of the group models. After discussing the groups’ model development in the first model development sequence, I summarize each focus groups’ model development. I use these summaries to exhibit how the focus groups’ models developed as they progressed through the first model development sequence, using repeated sampling, and the reasoning constructed by the groups before moving to the bootstrapping activities in the second model development sequence. I revisit these summaries after detailing the groups’ model development over the second model development sequence in order to discuss how groups’ reasoning constructed in the first model development sequence was used to construct reasoning in the second model development sequence.

My second research question addresses how students developed their reasoning of bootstrapping methods in order to make informal inferences about a population of data. While this question only addresses bootstrapping methods, the groups’ models constructed via repeated sampling activities built a foundation for the groups’ models of resampling and must be discussed first. I discuss this research question throughout my discussion of groups’ model development during the second model development sequence, and summarize the results at the end of the chapter.
My final research question, regarding student reasoning revealed and supported by moving from using hands-on manipulatives to computer simulations during repeated sampling and bootstrapping activities, is detailed throughout the discussion of groups’ model development in both model development sequences. At the end of the chapter, I bring together results related to this research question and focus on the role of the hands-on manipulatives and technology on developing the groups’ models of sampling and inference.

**Pre/Post-Test Results**

While the participants in the study come from both the high school and community college settings, I do not make a distinction between the developing models of sampling and inference based on the groups’ setting. Through my analysis of the pre-test data, I found no significant evidence supporting the lack of homogeneity of high school and community college introductory statistics students’ backgrounds in statistics and inferential reasoning (Table 2). This homogeneity of the settings allowed me to analyze the results of each setting as one. Further, Cohen’s effect size value ($d = 0.36$) suggested low to moderate practical significance.

Table 4

*Summary Statistics and One-Way ANOVA Analysis of Pre-Test Scores between the Two Settings*

<table>
<thead>
<tr>
<th>Groups</th>
<th>Count</th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>High School Participants</td>
<td>32</td>
<td>4.625</td>
<td>2.887</td>
</tr>
<tr>
<td>Community College Participants</td>
<td>33</td>
<td>4.000</td>
<td>3.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Settings</td>
<td>6.346</td>
<td>1</td>
<td>6.346</td>
<td>2.155</td>
<td>0.147</td>
</tr>
<tr>
<td>Within Settings</td>
<td>185.5</td>
<td>63</td>
<td>2.944</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>191.846</td>
<td>64</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

There was significant evidence to suggest gains in statistical reasoning over the course of
the instructional when examining all participants’ pre/post-test scores (Table 3). When I compared the pre/post tests for each setting and then each class individually, the community college students, and Brenda’s class with focus group D were the only setting and class to show significant gains in statistical reasoning. While there is statistically significant evidence of gains in the scores for these two groups, the gain is not likely of practical significance. Statistically significant evidence was shown by groups with mean increases in pre/post-test scores ranging from 0.557 to one point on the 13 item pre/post-test. Since the increase in scores represent a small increase proportionally to the total number of items, participation in the two model development sequences is likely not a practical means for increasing students’ statistical reasoning as examined by the pre/post-test.

Table 5

*Summary Statistics and Paired t-test for Increase in Pre/Post-Test Scores*

<table>
<thead>
<tr>
<th>Groups of Participants</th>
<th>Count</th>
<th>Mean</th>
<th>Variance</th>
<th>t</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>61</td>
<td>0.557</td>
<td>4.851</td>
<td>1.977</td>
<td>0.026*</td>
</tr>
<tr>
<td>High School</td>
<td>32</td>
<td>0.375</td>
<td>6.435</td>
<td>0.836</td>
<td>0.205</td>
</tr>
<tr>
<td>Community College</td>
<td>29</td>
<td>0.759</td>
<td>3.190</td>
<td>2.287</td>
<td>0.015*</td>
</tr>
<tr>
<td>High School Class One</td>
<td>19</td>
<td>0.000</td>
<td>6.000</td>
<td>0.000</td>
<td>0.500</td>
</tr>
<tr>
<td>High School Class Two</td>
<td>13</td>
<td>0.923</td>
<td>7.077</td>
<td>1.251</td>
<td>0.117</td>
</tr>
<tr>
<td>Community College Class One</td>
<td>17</td>
<td>0.588</td>
<td>3.382</td>
<td>1.319</td>
<td>0.103</td>
</tr>
<tr>
<td>Community College Class Two</td>
<td>12</td>
<td>1.000</td>
<td>3.091</td>
<td>1.970</td>
<td>0.037*</td>
</tr>
</tbody>
</table>

*Significant at the 0.05 significance level

NOTE: Only participants completing both the pre-rest and post-test are included in table.

Models of Sampling and Inference

The first model development sequence focused on repeated sampling from an available population and the drawing of inferences from empirical sampling distributions (see Figure 4). The goal of the sequence was to elicit and develop models of sampling and inference that the
students could use to simulate data and draw conclusions about the likeliness of possible outcomes based on the simulated data.

**Initial models of sampling and inference.** The model development sequence began with a model eliciting activity that elicited a method of simulating data and observing the patterns in the simulated data to draw inferences. The goal of the first MEA was to elicit a model that students could use to generate data and to examine the relative frequency of the outcomes in the data to determine which outcomes are most likely to occur. The MEA (see Appendix C) told a brief story of Paul the Octopus who had predicted eight of eight soccer matches correctly during the 2010 World Cup. Local zookeepers planned to have their octopus, Ophelia, predict the winners of eight upcoming basketball games. Groups of students were asked to write a letter to the newspaper with the range of outcomes that they thought were likely for the Ophelia the Octopus to correctly predict, including a description of the methods that they used to come to this conclusion. The students were also given a cup of 10 coins to elicit the idea that the octopus was guessing the outcome of each basketball game. These guesses could be simulated by flipping a coin. Students could simulate data for Ophelia’s predictions for the eight basketballs games by repeatedly flipping eight coins and observing the frequencies of various outcomes. From the frequencies, students could determine which outcomes were more or less likely to occur.

Two broad categories of models for sampling and inference were elicited from groups of students when engaged in this MEA: simulation models and calculation models. Groups that constructed simulation models (n=15, including focus groups A, B, C, and D), recognized that guessing the winner of a basketball game could be simulated by flipping one coin. Ophelia’s chances of correctly predicting the winners of eight basketball games could then be simulated by flipping eight coins. Groups that constructed calculation models (n=5) did not use the cup of
coins. Models were elicited that the students used to calculate the probabilities for Ophelia to predict various numbers of games correctly. Groups in this category constructed three different models that were used to calculate probabilities, with the differences being what the groups claimed were the total number of ways that Ophelia could predict the eight basketball games (16, 64, or 256).

Simulation models. Groups of students (n=15, including focus groups A, B, C, and D) used eight coins to simulate predictions for Ophelia. The groups recognized that an outcome that has a 50-50 chance of occurring could be simulated by flipping a fair coin. Initially, two types of simulation models emerged and differed by what the sides of the coins represented during simulation. Some groups (n=8, including focus group A) asserted that the sides of the coin represented Ophelia’s correct and incorrect predictions. A non-focus group discussed their simulation by stating that they tossed “8 coins multiple times. We assigned heads for correct and tails for incorrect. After the tosses we concentrated on the number of time[s] the coins landed heads side up” [Written Classwork, Ophelia the Octopus, Day 1]. This group focused on the number of Ophelia’s correct predictions that they represented as the head of each coin. Other groups (n=7, including focus groups B, C, and D) claimed that the sides of the coins represented the outcomes of winning or losing a basketball game. Focus group D stated, “To simulate [local] University’s next 8 basketball games we took 8 pennies and flipped each coin as it being one game. Heads represented a win and tails represented a loss” [Written Classwork, Ophelia the Octopus, Day 1].

These two initial models, incorrect vs. correct and win vs. loss, demonstrated that groups had alternative interpretations for the situation presented in the MEA. For the situation stated in the MEA, the groups were asked to determine how many outcomes Ophelia would correctly
predict. It did not matter which team won, only that Ophelia predicted the winning team. These competing models were briefly discussed during whole class discussion and group discussions with the instructors where the instructors briefly asserted the correctness of the first model of incorrect vs. correct. In Brenda’s class, the exchange below took place during a whole class discussion while a group was presenting their approach to the MEA that used the model of incorrect versus correct.

Zed: So it’s not about the games winning or losing, it’s about like, how does she predict what’s right?

Brenda: Let’s talk about not win or lose, but we’re talking about predictions, right?

[Class discussion, Ophelia the Octopus, Day 1]

Brenda asserted that the coins represented Ophelia’s predictions rather than the outcomes of the basketball games. This same assertion occurred in each of Allen’s classes. This was not an ideal situation for the elicitation of groups’ models. The self-evaluation principle for model eliciting activities states that the activity should allow the students to see the need to refine their models, while in this case, they were directed by the instructors to use the sides of the coins to represent Ophelia’s correct or incorrect predictions. This was both instructors’ first experience teaching with a modeling approach and my first experience working with instructors to implement model development sequences, so it was not surprising to observe a situation such as this. I did not adequately address this type of situation during my planning sessions with the instructors.

After simulating Ophelia’s predictions with the coins, three subcategories of simulation models emerged to determine the likely range for Ophelia’s correct predictions: 1) no conclusions were drawn from the data (n=1, including focus group C); 2) the likely range of correct predictions was constructed using all observed outcomes (n=12, including focus groups B
and D); and 3) an unlikely range of correct predictions was constructed containing numbers of correct predictions greater than all observed outcomes (n=2, including focus group A).

Focus group C produced the first subcategory of the model. They simulated three predictions by flipping the eight coins three times. They found differing numbers of correct predictions in each set of flips (3/8, 4/8, and 5/8 correct). Focus group C concluded that because of varying values in their simulations, there was “no definitive answer as to the range of possible outcomes” [Written classwork, Ophelia the Octopus, Day 1]. The group wanted to determine a single value for a correct answer and the variability in the collected samples did not appear to allow them to do so.

The second subcategory of model was constructed by the majority of the groups (n=12, including focus groups B and D). These groups produced a likely range of correct predictions that directly coincided with the outcomes that were simulated while flipping the coins. Focus group D flipped the set of eight coins 10 times and observed the outcomes of two, three, four, five, and six correct predictions. The group created a visual representation of this data (Figure 10) and drew a bracket around all observed to denote their range of likely outcomes. Focus group B followed the same approach and asserted that all simulated predictions constituted the likely range for Ophelia’s correct predictions.
Focus group A claimed that the outcomes that they observed were common for guessing the winner of eight basketball games. If Ophelia did better than the observed outcomes, the group claimed that she may have some ability to predict games.

*Calculation models.* All groups in this category asserted that Ophelia has a $\frac{1}{2}$ chance of predicting the outcome of each basketball game. The groups used this probability, along with the eight games that Ophelia was to predict, to construct three different models to determine the probabilities of predicting various numbers of games correctly.
In the first model, groups (n=3) determined that there were 16 different ways that Ophelia could predict the winners of eight basketball games. One non-focus group claimed that there were 16 possible outcomes, since there were two possible outcomes for each prediction, right or wrong, and eight basketball games to predict. They came to the conclusion that the “most probable” outcome was for Ophelia to predict four games correctly and as you moved further away from four correct predictions in either direction, the chances of that outcome occurring decreased (Figure 11).

Out of the 8 total games being played by [local university], Ophelia has a ½ or 50% chance of predicting each game correctly … There are 16 total outcomes – factoring in a 50% chance for each game. A breakdown of the probability range is below. [Written classwork, Ophelia the Octopus, Day 1]

![Figure 11. Non-focus groups’ table for the probabilities of guessing the number of correct predictions.](image)

Each group that claimed that there were 16 possible ways that Ophelia could predict the eight games found probabilities equivalent to those in Figure 11. No group provided a rationale in either their written work or class presentation for how these probabilities were calculated. One groups determined that “…the probabilities become greater towards the middle…”, which
demonstrated an understanding of the shape of the distribution. These groups correctly identified aspects of the shape of the distribution, such as the peak of the distribution at four correct predictions and decreasing probabilities in each direction.

One group claimed that there were 64 possible outcomes for Ophelia’s predictions, but did not provide any rationale for this number in their written classwork or during the class presentation of their model. This non-focus group initially focused on the probability of one outcome, when Ophelia predicted all basketball games correctly. The group stated that “\((1/2)^8 = 0.00390625\) is the probability of getting all matches right” [Written classwork, Ophelia the Octopus, Day 1]. Since the probability of this outcome equals 1/256, it appears to contrast with the group’s claim of only 64 outcomes. The group then produced a range of likely outcomes for Ophelia’s predictions in terms of their 64 possible outcomes and asserted that, “There are 64 outcomes possible, and the most likely are between the 16 and 48 outcomes…. The outcomes along the median of the predictions will be more likely” [Written Classwork, Ophelia the Octopus, Day 1]. The group constructed a model to determine likely versus unlikely outcomes, by noting that the 32nd, or the middle outcome, was the most likely outcome to occur and those outcomes moving away in each direction were then decreasingly likely.

The remaining group in this category of model concluded that there were 256 possible outcomes for Ophelia’s predictions. The group started to construct a tree diagram with two branches every split and \(1/2\) written on each branch (Figure 12). Tree diagrams had recently been used in the community college classes’ coursework. The group completed the tree diagram for Ophelia’s prediction of three basketball games rather than all eight games. This was likely because the size of the tree diagram quickly grew large when drawing new branches for Ophelia’s predictions. From the first few branches the group may have determined that to predict
each game correctly Ophelia would travel along eight branches of the tree diagram.

*Figure 12. Non-focus group’s tree diagram for Ophelia’s eight predictions.* [Written classwork, Ophelia the Octopus, Day 1]

The group determined that the probability that Ophelia would predict all eight games correctly is \((1/2)^8 = 1/256\). This non-focus group described during the whole class discussion how they came to this conclusion.

Brenda: Back up to the one over 256. Where does that come from?

Xander: It’s \(\frac{1}{2}\) times \(\frac{1}{2}\), times \(\frac{1}{2}\), eight times.

Zach: That’s like the chance of her guessing each individual game correctly. [Class discussion, Ophelia the Octopus, Day 1]

This group focused on the probability that Ophelia predicted all eight games correctly, claiming that it was unlikely for her to do so. They determined, “It is very unlikely for her to predict every single one correctly, only having a .004 probability of doing so while it is much more likely for her to predict one game correctly” [Written Classwork, Ophelia the Octopus, Day 1] The focus was on the unlikeliest event, rather than a range of likely events.
Using models to evaluate past outcomes. While the MEA aimed to predict future outcomes, the homework assignment following the activity asked individual students to use their reasoning from the MEA to evaluate the likeliness of an outcome that had already occurred. The homework assignment discussed Paul the Octopus’s eight correct predictions during the 2010 World Cup and was completed by 48 individual students. The students were told that a blog claimed that Paul must be psychic and asked to comment on the blog post (Appendix C). All responses made some remark about it being unlikely for Paul to correctly guess the winners of eight matches. The differences occurred in how this unlikeliness was interpreted.

Some students (n=11 students) determined that since the outcome was unlikely, Paul must not be guessing the outcomes, but instead be psychic. Only students from the high school (all of whom used the method of simulations with coins in the MEA) drew this conclusion. Cara, from focus group A, concluded that:

There is fairly convincing evidence that Paul is psychic. This is because I simulated ten trials of coin flipping; flipping it eight times each trial … In my results, I did not get higher than 5/8 correct. For me to have convincing evidence of a special ability, Paul would have had to predict more than 5 of the outcomes correctly. Considering that he got all 8 correct, I would look into that psychic theory. [Written homework, Paul the Octopus Controversy, Day 1]

Cara compared Paul’s eight correct predictions to her range of observed values for the number of Ophelia’s correct predictions. Since Paul’s predictions were outside of this range, she concluded that it would be unlikely for Paul to guess the eight games correctly and he must have a “special talent”. I interpret this language to indicate that Cara has concluded that Paul is not guessing the outcomes and therefore has a better than 50-50 chance of predicting the winner of a soccer
The remaining students (n=37 students) concluded that although it is very unlikely for an octopus to predict the winners of all eight matches, it is a possible outcome. Students who used either methods of calculating probabilities or simulations with coins came to this conclusion. Megan (focus group C) concluded that there was only a \(\frac{1}{256}\) chance of Paul predicting all eight winners “but it is still a possibility. So just because this very improbable thing happened doesn’t make an octopus psychic.” Megan asserted that all outcomes are possible; just because the unlikely occurred, this does not imply that Paul is psychic. Megan decided to read ahead in her statistics textbook to find a “better” way to calculate a likely range of values for Paul’s correct predictions (Megan, Focus Group C, Interview 1). Megan calculated the normal approximation to the binomial distribution and determined that any outcome more than two standard deviations from four correct predictions would be considered unlikely (Figure 13).

![Image](https://example.com/image.png)

*Figure 13. Megan’s solution to Paul the Octopus Controversy. [Written homework, Paul the Octopus Controversy, Day 1]*

For this homework, individual students used their reasoning from the MEA or other prior knowledge to make sense of unlikely events. Some students thought that guessing eight out of eight matches was so unlikely that Paul must be psychic, while others acknowledged that it was
unlikely to have happened by chance, but it was still possible, and more reasonable than claiming that an octopus was psychic.

Impact of technology on developing models. The next activity in model development sequence was a model exploration activity (MXA) that was meant to examine the structure and representations of the models elicited in the MEA. The goal of the MXA was to explore the dotplot representation of increasing numbers of simulated data and the structure of the sampling distribution. The MXA used the same context as MEA, but students now used TinkerPlots (Konold & Miller, 2014) to simulate outcomes rather than using the coins (Appendix C). The MXA asked the students to compare the TinkerPlots environment to the methods that groups used in the MEA. Several non-focus group students noted that “it’s similar because the tinkerplots represent[s] the spinner to have a 50/50 chance just like a coin would have a 50/50 chance” and “we had 8 coins and with Tinkerplots we could use the spinner 8 times” [Written classwork, Ophelia with TinkerPlots, Day 2]. These students demonstrated an understanding of the similarity between simulating Ophelia’s outcomes with eight coins, and the use of the spinner in TinkerPlots. The activity guided the students to explore how the collection of more trials of Ophelia’s predictions affected the conclusions that they drew from the data. Each student individually used the TinkerPlots environment to simulate 10, 50, and 1000 trials and determined the likely range of correct predictions by Ophelia for each number of trials. Within each group of students, they compared the similarities and differences between each student’s ranges for each number of samples.

Two developments occurred in groups’ models of sampling and inference in this MXA. First, some groups focused on the effect of collecting more trials and found that increasing the number of trials increased the similarity between the predictions of the groups’ members. Other
groups concluded that regardless of the number of simulated trials, each student’s range of likely values for Ophelia’s predictions was similar. Second, some groups of students no longer claimed that all simulated outcomes were likely outcomes. As the number of trials increased, many students observed outcomes occurring infrequently that they had thought were unlikely. Groups now needed to determine a method for concluding which observed outcomes were likely versus unlikely.

The effect of more trials. Groups noted that the outcomes shown in each of their dotplots after simulating 10 trials were not the same among group members. In focus groups A and D, each group member’s predictions were similar, but slightly different. For the three group members in focus group A, the outcomes fell between two to five, three to five, and two to six correct predictions (Figure 14). Cara noted that she “thought that maybe 4 would have the most dots, but 5 did” [Written classwork, Ophelia with TinkerPlots, Day 2]
Figure 14. Focus group A’s dotplots for each student’s 10 sets of Ophelia’s predictions

Cara used the three sets of 10 samples shown in Figure 14 to determine that “it’d be surprising if
Ophelia predicted less than 2 correct or more than 6” [Written classwork, Ophelia with
TinkerPlots, Day 2].

Only two of the three members of focus group B (George and Henry) were present for
this activity and their results were quite different from each other. Most of Henry’s outcomes
were as he expected them to be, in the range of four to six correct predictions, but he also
generated outcomes of one and two. Henry’s partner George stated that these outcomes “seemed
very strange because they are so far off from what the outcomes should have been (4/8, 50%)”
[Written classwork, Ophelia with TinkerPlots, Day 2]. George had similar results with most of
his values between three and five correct predictions with additional outcomes of one and seven
that he also thought seemed “very unusual because it is so far from 50%” [Written classwork,
Ophelia with *TinkerPlots*, Day 2]. Students in focus groups A and B expected most of their outcomes to be in the middle of the dotplot since it was most likely for Ophelia to guess half of the games correctly. Cara from focus group A claimed “there would be less variation in a smaller sample size” [Written classwork, Ophelia with *TinkerPlots*, Day 2]. In this context, I interpreted this as meaning for a smaller number of simulated trials. Cara thought that since the outcomes with lower and higher numbers of correct predictions were unlikely to occur, she would need to simulate many trials in order to have these outcomes occur by chance.

As focus group B moved to taking 50 trials, George and Henry noted that each other’s dotplots were “very similar and all/most of the dots were around 50% or 4/8” [Written classwork, Ophelia with *TinkerPlots*, Day 2]. Henry noted that each of their dotplots had a “pyramid shape” [Written classwork, Ophelia with *TinkerPlots*, Day 2]. They concluded that the likely ranges were three to six correct predictions and three to five correct predictions. After simulating 1000 trials, Henry noted that they had “identical graphs” [Written classwork, Ophelia with *TinkerPlots*, Day 2]. George stated that they were “very similar” in “that all dots / number of dots per # of games correct were very similar” [Written classwork, Ophelia with *TinkerPlots*, Day 2]. I interpret this to mean that he was referring to the number of times that each outcome occurred was similar between him and his partner’s dotplots. Both students concluded from these dotplots that Ophelia is likely to predict three to six of the eight basketball games correctly.

Focus group C, along with some other non-focus groups, did not discuss how differing numbers of samples led to changes in each students’ interval of likely ranges of correct predictions, or in group differences in these intervals. Megan stated that when her group members each simulated 10 trials, two of their dotplots:

… had some form of a bell curve, while one was skewed to the left …They all thought it
was going to be a bell curve, but the results didn’t comply. They [the dotplots] should be the same because all trials are similar and they are independent and there is a 50/50 chance of it being correct or not. [Written classwork, Ophelia the Octopus with *TinkerPlots*, Day 2]

Megan claimed that the “results didn’t comply”. This is similar to her methods used in the homework of wanting to use calculations to answer the activities rather than simulation. As focus group C collected 50 and then 1000 trials, they claimed that each plot looked like a bell curve and did not change their predictions for the likely ranges of correct predictions. Lena (focus group C) claimed, “the more trials conducted will help conclude a more accurate prediction” [Written classwork, Ophelia with *TinkerPlots*, Day 2]. I interpret that by more accurate, Lena means that as more trials were simulated, the dotplot would look more like the bell curve that the group members had anticipated. Megan continued to add trials and came to the conclusion that at some point, adding more trials did not change the look of the dotplot. She concluded that the plot was “saturated with data” [Written classwork, Ophelia with *TinkerPlots*, Day 2].

**Determining ranges of likely outcomes.** When simulating 10 trials of Ophelia’s predictions, focus group B did not make a prediction for Ophelia’s likely range of correct predictions. As they stated earlier, they had many values that they considered unusual, with George stating that his dots are “scattered everywhere” [Group discussion, Ophelia with *TinkerPlots*, Day 2]. When simulating 50 trials, focus group B no longer used all observed outcomes as the likely range of values. This group considered the outcomes that they deemed occurred the most often. Outcomes ranging from one to seven correct predictions occurred in each of their collections of 50 trials, a wider range than when simulating 10 trials. Using their dotplots (Figures 15 and 16), Henry concluded that the likely range for correct predictions was
three to six, while George claimed three to six correct predictions. In each case, they made a decision of which values occurred enough to be likely and which values occurred too few times and were considered unlikely by examining the dotplot created in the TinkerPlots environment. They did not articulate a specific rule for determining enough versus not enough, and referenced the number of dots for each number of correct predictions. Henry’s predicted range was larger than George’s because he could not rule out the value of six as being likely as done by George. In Henry’s simulations, five correct predictions only occurred one more time than six correct predictions. This was very different in George’s dotplot, where five occurred eight more times than six.

Figure 15. Henry’s dotplot for 50 trials of Ophelia’s predictions (focus group B).
Developing models of sampling and inference. A homework assignment completed by individual students (n=25 students) returned to the same context as the MEA with Ophelia the octopus. All students were asked to complete the assignment, but less than half of the participants turned in the completed assignment. Individual students were asked to write another letter to the local newspaper with the range of outcomes that they thought were likely for Ophelia to correctly predict. The students were told that they could use any of the data that they had gathered with the cup of coins or TinkerPlots in their analysis (Appendix C). Most students used similar methods as observed in the MEA and MXA, such as a focus on one outcome rather than a range (n=3 students), a predicted range that contained all observed values (n=1 student), and a predicted range containing most of the outcomes (n=19 students). A development observed for some students (n=2 students) was an extension of the idea of the range containing most values, using calculations of proportions to justify the choice of ranges.

Focus on one outcome. Students in this category (n=3 students), including those in focus group B, focused on the outcome of four correct predictions. Henry stated that “Based on Tinkerplots it shows that 50% (4/8) outcomes most of the time … out of eight games Ophelia
should choose 4 right” [Written homework, Ophelia Take Two, Day 2]. These students focused on the expected value of the number of correct predictions because it was the most occurring outcome.

All observed values. In the MEA, many students constructed their ranges of likely to contain all outcomes that occurred when simulating a small number of trials. Since the TinkerPlots activity guided the students to simulate 1000 trials, it was likely that all possible outcomes would occur. Only one student (non-focus group) continued to use this method of inference and stated that “the range of outcomes can be from all eight lost to all 8 won” [Written homework, Ophelia Take Two, Day 2]. He did not use any language for likely or most common outcomes and asserted that all outcomes were possible.

Most values. The majority of students (n=19 students) reasoned about which outcomes occurred more often than others to determine the likely range of correct predictions. A non-focus group student stated:

Using the data from Tinkerplots, I’ve come to find out that it’s very likely for Ophelia to predict 3-5 games correctly. It’s somewhat less likely for her to guess 2 or 6 games correctly, It’s much less likely for her to guess 1 or 7 games and it’s almost impossible for her to correctly guess 0 or 8 games correctly, but there’s still a chance. [Written homework, Ophelia the Octopus Take Two, Day 2]

This student demonstrated an understanding of the symmetry of the likelihoods of the outcomes and that outcomes that occurred fewer times in his sampling were less likely to occur for Ophelia’s predictions. Much like the previous MXA, students in this category did not provide a rationale for the dividing line between unlikely and likely other than an event occurred more. For instance, the student above did not elaborate why he thought that three to five correct predictions
were likely rather than two to six other than three and five occurring more than two and six.

Students in focus group C followed a similar line of reasoning to determine that it would be likely for Ophelia to predict two to seven games correctly. Megan stated that:

…the mean looks like it is 4 and the next 2 # of right predictions is 3 and 5 with about 200 correct matches. 2 right predictions and 7 right predictions have 100 correct predictions. Therefore w/ this information I would say the mean is 4 and the range is 2 – 7. [Written homework, Ophelia the Octopus Take Two, Day 2]

Megan appears to have misstated a few things in her statement. She first stated that “7 right predictions have 100 correct predictions” rather than stating that the outcome of seven correct prediction occurred 100 times. Her conclusion also contained a dotplot of her outcomes with 1000 simulations and 6 correct predictions occurred 100 times, not seven correct predictions. She may have misread the dotplot, which resulted in a range of predicted values that was not symmetric around her concluded mean value.

Other students in this category, including students in focus group A, used the representations of the simulated data from TinkerPlots to update the ranges of correct predictions that they constructed in the MEA. Cara stated that:

In my previous investigation, I collected data by flipping coins …. Seeing only 10 trials of the coin flipping method, I came up with the likely range of outcomes to be 2-5/8 correct. After using the TinkerPlots software, where I could view up to 1000 trials, I would think that getting 6/8 correct would also be a likely outcome. [Written homework, Ophelia the Octopus Take Two, Day 2]

With observing a small number of outcomes, Cara did not see that two and six correct predictions were equally likely, but with 1000 trials, this was much clearer. Her work with the
MXA revealed to Cara the symmetric structure of the empirical sampling distribution.

Proportions to justify ranges. Two students used proportional reasoning to evaluate the likely outcomes for Ophelia’s predictions. A non-focus group student used TinkerPlots to simulate 13,335 trials. He then used the dotplot to approximate the number of trials for each outcome (Figure 17).

\[ \begin{array}{cccccccc}
0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
5 & 285 & 2,750 & 3,500 & 3,500 & 2,750 & 1,500 & 550 & 5
\end{array} \]

Figure 17. Non-focus group student’s table for the number of trials collected for numbers of correct predictions.

The student stated, “by dividing the number of each individual category by the total number of trials, I can calculate the chance of each outcome somewhat accurately” [Written homework, Ophelia the Octopus Take Two, Day 2]. He claimed that probability of each outcome could be determined by looking at the proportion of the number of times the outcome occurred, divided by the total number of samples. The student did not then use the table or perform calculations to determine likely versus unlikely outcomes.

Randy, a student in focus group D, determined his likely range of correct predictions by taking into account the proportion of the samples that fell into his range of values. In his original predicted range of outcomes, Randy determined that Ophelia was likely to predict two to six games correctly. After using TinkerPlots to collect 1000 trials, Randy stated, “I found that 74% of the results fell into the range of 3 right, 5 wrong to 5 right, 3 wrong with the remaining results
falling outside of this range” [Written homework, Ophelia Take Two, Day 2]. This proportional thinking parallels the thinking used in formal inference for a confidence interval. Randy decided that 74% of the samples constituted a large enough proportion of the total samples. Randy informally constructed a form of a 74% confidence interval.

**Application of models of sampling and inference.** In the MEA, students simulated (or calculated) Ophelia’s predictions by assuming the probability of guessing the winner of a game was equal to \( \frac{1}{2} \). In this MAA, groups of students had to determine the probability of each outcome in a new context. The goal of the MAA was to apply the model to support or oppose claims regarding the population from which the data was drawn using simulated data as evidence. Students were told that in 2013, 35% of all sneakers sold globally were Nikes (Appendix C). A marketing director from Nike contacted the students to help him determine if this global trend of Nike sneaker sales held true for the students at a local university. The marketing director planned to discuss sneaker sales for [local] University students at an upcoming meeting and asked the groups of students if it was reasonable to claim in his meeting that about seven in 20 [local] University student sneaker purchases are Nikes. The groups of students were given a bin containing thousands of multicolored beads and told that each bead in the bin was a sneaker purchase by a student from the university and every clear bead was a Nike sneaker purchase.

All groups collected between 10 and 20 samples of 20 beads from the bin and then used these samples as evidence to make claims about the university students’ sneaker purchases. Three categories of models were constructed, differing in how the sample data was used to support or reject the claim of 35% of sneaker purchases at the university being Nikes. Groups in the first category (n=7, including focus groups D) used only measures of center for the number
of Nike sneaker purchases in their samples as evidence to make claims about the percentage of Nike sneaker purchases made by students at the university. Groups in the second category (n=2) used only the distribution of the samples as evidence for claims about the percentage of Nike sneaker purchases made by students at the university. Groups in the third category (n=10, including focus groups A, B, and C) used both measures of center and the distribution of the samples as evidence for claims about the percentage of Nike sneaker purchases made by students at the university.

Center value as evidence. Groups constructing this category of model found measures of center for the Nike sneaker purchases in their collected samples. The groups observed the center of a dotplot of collected samples or calculated the mean number (or percentage) of Nike sneaker purchases in their samples, which were used as evidence to predict the Nike sneaker purchases at the university. Focus group D collected 10 samples of 20 sneaker purchases and plotted the number of Nikes in each sample on a dotplot (Figure 18).

![Figure 18. Focus group D’s dotplot of Nike sneaker purchases](Written classwork, Nike Market Share, Day 3]

The group discussed the distribution of their outcomes by talking about how most outcomes (8, 10, and 12) were grouped together, with only two of the outcomes (13 and 14) slightly outside of this group. Although the distribution of samples was briefly discussed, the group did not use it to draw their conclusion for their written classwork. Focus group D found the mean percentage of
Nike sneaker purchases in their samples to be 55% and then used the sample mean as the predictor for the population of student Nike sneaker purchases without referencing the Nike marketing director’s claimed value of 35%. Focus group D asserted that 55% of sneakers purchased by students at the local university were Nike sneakers. Focus group D stated:

Based upon the total of 109 purchases of Nike’s out of 200 total student purchases we come to the conclusion that 55% of students purchase Nike sneaker based upon data.

[Nike Market Share, Written classwork, Day 3]

Focus group D’s conclusion combined the samples into one group of 200 observations which condensed the data and did not show aspects of the data, such as the variability between samples, which was used by other groups as evidence to dispute the Nike marketing director’s claim.

A non-focus group also constructed a model using the mean of their samples as evidence. This non-focus group calculated a mean of 9.5 Nike sneakers purchased for every 20 by the university students and determined that this mean was close to the claimed value and therefore supported the marketing director’s claims. They concluded, “Out of the 20 trials we came up with an average of 9.5/20 clear beads. The claim 7/20 people purchasing Nike shoes is reasonable” [Nike Market Share, Written classwork, Day 3]. This non-focus group did not indicate what mean values of Nike sneaker purchases would be close enough to support the claim of 7/20 Nike sneaker purchases or far enough to dispute that claim. This is in contrast to focus group D, who did not compare their sample mean to 7/20, and only predicted that their sample mean was equivalent to the population mean.

*Distribution as evidence.* In Allen’s class with focus group A, the samples from all groups were aggregated onto the interactive white board in front of the classroom (Figure 19).

A non-focus group determined that seven Nikes out of 20 sneaker purchases was not likely since
most of the simulated trials were greater than the claimed value. This group considered the proportion of outcomes greater than 7/20. This non-focus group determined that “It is reasonable to up the initial claim of amount of 7/20 because 74 of the 105 trials were greater than 7/20” [Nike Market Share, Written classwork, Day 3]. Since a large proportion of the trials were greater than the claimed value, the group asserted that this was evidence against the Nike Marketing Director’s claim. A different non-focus group from the same class used the aggregated data to determine, “Anything below 7 and above 13 is not very likely. There was 15/105 below 7 and 6/105 above 13 which is a low number” [Nike Market Share, Written classwork, Day 3]. Evidence citing the proportion of outcomes greater (or less) than a given value is a step towards formal inferential reasoning.

![Figure 19. Aggregate class data of 105 samples of 20 sneaker purchases (focus group A’s class).](Class discussion, Nike Market Share, Day 3)

**Center and distribution as evidence.** Focus group C initially collected 20 samples of 20 sneaker purchases and constructed a dotplot of the data (Figure 20). From this dotplot, the group determined that 10 or 11 Nike sneaker purchases was the center of their collected values. The
Nike marketing director’s claim of seven of 20 sneaker purchases being Nikes was asserted to appear to be low.

Figure 20. Focus group C’s dotplot of Nike sneaker purchases. [Written classwork, Nike Market Share, Day 3]

Nate: He [Nike marketing director] was saying that about 7 out of the 20 people were going to buy Nike sneakers, so, from what we got there looks like there’s going to be more.

Megan: Yeah, it looks like from our data it would be more like 10 or 11, because if we kept doing it I’m guessing that 10 or 11 would be like the top of it. [Group Discussion, Nike Market Share, Day 3]

Megan indicated that if they collected many more samples the peak of the dotplot (Figure 20) would occur at either 10 or 11 out of 20 Nike sneaker purchases. At this point they had not calculated any measures of center and only compared their dotplot to the claimed value of seven of 20 sneaker purchases were Nikes. The outcomes of 10 and 11 Nike sneaker purchases occurred fewer times in focus group C’s collected samples than outcomes such as nine and 12. Megan still asserted that if more samples were taken, they would be the most likely to occur since those outcomes would be in the center of the dotplot. Megan anticipated the shape of the
distribution and determined where the center of the distribution would be.

Focus group C then used the center of their distribution of samples of sneaker purchases to determine a likely range for the number of Nike sneaker purchases out of 20 sneaker purchases. In their letter to the Nike marketing director focus group C concluded that: “Based on our data saying about 7 in 20 students who purchase Nike would be wrong. The mean seems to be 10-11, and the range would be 9-12.” [Nike Market Share, Written classwork, Day 3]. Focus group C determined a likely range of outcomes for the number of Nike sneaker purchases out of every 20 from their empirical sampling distribution and compared it to 7/20 (Figure 20). The group excluded the three lowest and four highest values and asserted that the range of likely Nike sneaker purchases was made up of the middle 13 outcomes. Since the Nike marketing director’s claim was outside of this interval, the group asserted that it was incorrect. Focus group B drew their conclusion in a similar manner by constructing a likely range of values of eight to 11 Nike sneakers for 20 sneaker purchases. Since the value of 7/20 was below this interval, the group claimed that more than 7/20 sneaker purchases from the local university students were Nikes.

Focus group C later calculated the mean number of Nike sneaker purchases from their samples, 10.65, and compared this mean to the claimed 7/20. When presenting their conclusions to the class they did not discuss their dotplot and only compared the two values of 10.65 and seven Nikes out of 20 sneaker purchases.

Karl: I added up all the different, like, when we got 7 clear beads, 8 clear beads, I added all those up and divided by 20, and our average was 10.6

Brenda: 10.6?

Karl: 5, 10.65 was our average.

Brenda: So 10.65 white beads out of 20?
Karl: Yeah

Brenda: So tell me what the gist of your letter is to the Nike representative.

Nate: I said that seven out of 20 is probably a little low. In the 20 trials that we did, we got 10.6 around the average of Nike shoes that [university] students got.

[Group Discussion, Nike Market Share, Day 3]

Focus group C concluded that the claim of seven out of every 20 sneaker purchases at the local university being Nikes was low compared to their mean of 10.65 Nike purchases for every 20 sneaker purchases. The evidence that focus group C used during the class discussion was more in line with the first category of model, in which only the center of the distribution was used as evidence for inferential claims.

Focus group A initially collected 10 samples of 20 student sneaker purchases and constructed a dotplot of the outcomes (Figure 21). From this dotplot, Cara asserted, “well, based on our data, 7 out of 20 is low, because we got higher than that. So he can claim that it’s even higher” [Written Classwork, Nike Market Share, Day 3]. The outcome of 7/20 was the lowest outcome of sneaker purchases that they collected. Since the remaining outcomes were greater than 7/20, she determined that a higher number of Nike sneaker purchases was more likely. This initial reasoning only considered the distribution of samples and not the center of the distribution.

![Figure 21. Focus group A’s dotplot of Nike sneaker purchases](Group Presentation, Nike Market Share, Day 3)
Later focus group A’s class aggregated all samples of sneaker purchases from the class onto one dotplot (see Figure 19). The group used this sampling distribution to construct a likely range of four to 14 Nike sneaker purchases by constructing a range of values equidistant above and below the center of the distribution. Focus group A used this range of values to come to the same conclusion as focus group C, that the claim of 7/20 sneaker purchases being Nikes was too low, but in a different manner. Focus group C determined that 7/20 was outside of their likely range, hence unlikely. Focus group A determined that since 7/20 was not in the center of their range, it was not likely. Focus group A stated:

Based on this dot plot, we can infer that a likely range of outcomes for [local university] students’ Nike sneaker purchases would be 4-14/20, with a median of 9. I believe that 7/20 (35%) is a little low after viewing the results on the dotplot. [Written classwork, Nike Market Share, Day 3]

**Interpretations of the manipulatives.** Several groups of students had difficulty understanding how using the bin of beads differed from using the coins as hands-on manipulatives. In the MEA, students constructed models that assumed that Ophelia was guessing the winners of the basketball games, and knew that there was a 50-50 chance that flipping a coin would result in heads or tails. In the Nike sneakers MAA, the students were told that the global market share for Nike sneaker sales was 35% and they were to determine the likely range of Nike sneaker sales for the local university. This led some students to assert that 35% of the beads in the bin were clear beads, which represented Nike sneaker purchases. Focus group A stated that:

With 35% of all sneakers sold globally being Nike, I decided to use a bin of beads that contained three colors: orange, blue, and clear. The clear beads occupied 35% of the total
in the container, representing the company’s international sales performance. [Written classwork, Nike Market Share, Day 3]

This is in contrast to the results reported by this group with a dotplot that was centered around 9/20 Nike sneaker purchases after collecting 105 samples. Cara and I discussed the methods of sampling used in this MAA during an interview and how she would collect this data if she did not have the bin of beads.

Cara: I guess what you’d have to do is, say go to [local shoe store near the university] for a period of time … when someone purchases a pair of sneakers … you would have to record out of 20 purchases the amount of purchases that were Nike. That would take a very long time, so it’s good that we have the resources that we do… You’d have to get at least 100 trials, so like 2000 purchases. It would be a lot more time consuming, but it can be done. [Cara, Focus group A, Interview 1]

This method of sampling from the population is similar to how I had anticipated that the students would view the bin of beads. In this real-world scenario of sampling described by Cara, there is no assumption that 35% of the sneaker purchases are Nikes. When she simulated data with the bin of beads, she asserted that that assumption was needed in order to construct the composition of the beads in the bin.

A non-focus group also thought that to simulate the sneaker purchases, 35% of the beads in the bin should be Nike purchases. The group asserted that this could be approximated if there were the same amount of beads of each color, with 1/3 or 33% of the beads representing Nike sneaker purchases. The group stated that: “The odds of drawing a clear bead is one out three, similar to the 7/20 odds that the Nike director predicted” [Nike Market Share, Written classwork,
Day 3]. Neither Brenda nor I anticipated this view of the manipulatives (Brenda Instructor Debriefing, Day 3). This non-focus group did not discuss this thinking with the class during whole class discussion.

Focus group D also struggled with the proportions of each color of bead in the bin and how these proportions would affect their samples.

Randy: Are there an equal number of beads, each color bead? That would definitely make a difference on the outcome I would think. Wouldn’t it?

Ted: Yeah. [Brenda comes over to the group]

Randy: Do we know that there is an equal number of clear beads and the other color beads?

Brenda: We do not know that. What we know is that the beads represent … all sneaker purchases by [University] students. But that’s what you’re

Susan: …asking?

Randy: There may be more of one color?

Susan: So, what you’re saying is that we’re trying to figure out how many different shoes people will buy, and how many of those shoes are Nikes, and if this percentage (claim of 35%) is actually correct? [Group discussion, Nike Market Share, Day 3]

Randy may have been thinking back to the MEA with Ophelia, where it was equally likely for her to predict correctly or incorrectly the outcome of each basketball game. MAAs require adaptation of previous models. In contrast to the MEA, it is not known in the MAA if the chances for choosing different brands of sneaker purchases are equally likely. Susan draws on this at the end of the discussion to assert that the composition of the bead colors in the bin is
what they are trying to determine.

These interpretations of the bin of beads may have occurred because of the two different ways that the manipulatives were used in the MEA and this MAA. In the MEA, the eight coins were used as a tool to simulate sample predictions. In this MAA, the beads represented the population of sneaker purchases, which were used as a tool to simulate samples of sneaker purchases. I purposely designed the activities to use the manipulatives in both of these ways since both are key to the process of bootstrapping.

**Developing models with technology.** The next MXA stayed within the context of Nike sneaker purchases among the local university students and returned to using *TinkerPlots* to simulate data (Appendix C). The goals of the MXA was to explore the dotplot representations of simulated data to determine likely outcomes and compare of reason with simulated data generated from different original samples The sampler in the *TinkerPlots* environment contained the same number of balls as there were colored beads in the bin used in the previous MAA. The balls were marked with N (Nike), A (Adidas), and O (Other) to represent the three colors of beads in the bin. Students collected samples of 20 sneaker purchases in the *TinkerPlots* environment to investigate the claims that could be drawn from collecting varying numbers of samples. The MXA also guided the students to explore the shape of the distribution of the numbers of Nike sneaker purchases in their samples of 20 sneaker purchases, how these shapes varied between their group members’ dotplots, and how the shape of the distributions varied as more samples were collected.

At the end of the previous MAA, all groups of students in each class aggregated their samples of student sneaker purchases onto one dotplot at the front of their classroom. The number of total samples from the four classes were 120, 105, 55, and 47. The MXA asked the
students to simulate with *TinkerPlots* the same number of samples as they had done as a class and use the new data to make claims about the percentage of Nike sneaker purchases for students at the local university. Students were then asked if they were confident in the claims made from that number of collected samples and how many samples they would simulate to be confident in their claims and why. The two most common student responses were those that suggested collecting more than 1000 samples (n=21) and student responses that suggested collecting between 100-200 samples (n = 28). The reasoning given by students for choosing all numbers of samples was based on accuracy.

Students in focus group B all claimed that they would be confident in their claims if they collected between 100-200 total samples. George asserted, “100-150 trials to get the most accurate representation of what the most would be” [Written classwork, Nike Market Share with *TinkerPlots*, Day 4]. I interpreted that “most would be” is referring to the outcomes that occur most often compared to other outcomes. By collecting 100-150 samples, George appeared to think that the shape of the distribution would be clear enough for him to find the peak and center of the plot, or the outcome that occurs the most often. After collecting a total of 120 samples, Henry (also in focus group B) claimed that their dotplot (Figure 22) “has the same volcano look as the data in class. It comes up like it should but then goes down and then back up again. We are looking for the mountain look” [Written classwork, Nike Market Share with *TinkerPlots*, Day 4]. Henry was asserting that he wanted to collect enough samples for his empirical sampling distribution to look like a bell curve.
Students who wanted to collect 1000 or more samples used ideas of accuracy. Aaron (focus group A) claimed that “1000 trials because it is most accurate with a large sample size” [Written classwork, Nike Market Share with TinkerPlots, Day 4]. I interpreted that Aaron meant “large sample size” to be a large number of samples. Lena (focus group C) asserted, “By doing 1000 trials it will give us more accuracy” [Written classwork, Nike Market Share with TinkerPlots, Day 4]. More accuracy in these two students’ written work appears to also have the idea discussed by George of an empirical sampling distribution with the peak in the center and roughly “mountain” shaped [Written classwork, Nike Market Share with TinkerPlots, Day 4]. By collecting more samples, these students thought that the plots will be closer to this shape of distribution and hence more “accurate”.

Before collecting the number of samples that would make the students confident in their claims, the MXA asked the students to sketch a dotplot for what they thought that the dotplot of their collected samples of 20 sneaker purchases would look like. Lena (focus group C) sketched a curve (Figure 23) rather than a dotplot, and stated that she thought “it would look like an almost perfect bell curve” after collecting 1000 samples of 20 sneaker purchases [Written
classwork, Nike Market Share with *TinkerPlots*, Day 4]. Lena thought that after collecting 1000 samples, the dotplot would not deviate from this bell shaped pattern.

![Image](image_url)

*Figure 23.* Lena’s (focus group C) sketch for the number of Nike sneaker purchases for 1000 samples. [Written classwork, Nike Market Share with TinkerPlots, Day 4]

Students discussing accuracy may have been focusing on the variation between their dotplots as more samples were collected. When collecting increasing numbers of samples, the dotplots likely varied less from what the students viewed as the accurately shaped distribution. Randy (focus group D) went into more detail on this idea of accuracy and focused on the variation between his group members’ dotplots as more samples were collected. Randy claimed that number of samples taken should be “1500 because the more samples taken, the more likely possibility that our data results are closer to being the same” [Written classwork, Nike Market Share with *TinkerPlots*, Day 4]. By closer to being the same, Randy was going back to the thinking that his group used in the first MXA. In that activity, his group members compared each other’s dotplots as they collected 10, 50, and then 1000 samples. The group determined that their plots were increasingly more similar as more samples were collected. Randy was asserting that as different people collect large amounts of samples, the distribution of their data becomes more similar, which allows them to draw more similar conclusions from the data.

Some students based the collection of more samples on the likeliness of more samples producing a broader range of values. Non-focus group students stated, “I think that about 1000
samples should be simulated because there are more trials and more outcomes”, “149 should be ran to be confident so that you can get more data and a bigger range of values”, and “114 more times on tinkerplots would be fine because it gives you a little more range and more trials to see similarities” [Written classwork, Nike Market Share with TinkerPlots, Day 4]. In the first MXA with Ophelia, groups collected hundreds, or sometimes thousands, of samples before observing outcomes such as Ophelia correctly (or incorrectly) predicting all eight basketball games. In the Ophelia MEA and MXA there were nine possible outcomes, ranging from predicting zero to eight games correctly. In the Nike MAA and MXA, there were 21 possible outcomes, ranging from zero to twenty Nike sneaker purchases. Since there were many more outcomes, it was even more unlikely for students to collect the lowest and highest values in the range of possible outcomes. One way to likely cover more of the range of outcomes is to collect more samples. Veronica (non-focus group member) wrote the third quote above, stating that “114 more times on tinkerplots would be fine because it gives you a little more range and more trials to see similarities” [Written classwork, Nike Market Share with TinkerPlots, Day 4]. For the activity, Veronica first collected 55 samples (Figure 25) and then claimed that she would be confident in her claims for the predicted range of Nike sneaker purchases if she collected 114 more samples. In her initial 55 samples (Figure 24), the data ranged from three to 16 out of 20 Nike sneaker purchases. In her predicted dotplot after collecting 114 additional samples (Figure 25), she thought that the range would be larger, and spanned from two to 20 out of 20 Nike sneaker purchases.
Figure 24. Veronica’s (non-focus group) dotplot for the number of Nikes in samples of 20 sneaker purchases for 55 samples. [Written classwork, Nike Market Share with TinkerPlots, Day 4]

Figure 25. Veronica’s (non-focus group) predicted dotplot for the number of Nikes in samples of 20 sneaker purchases after collecting an additional 114 samples. [Written classwork, Nike Market Share with TinkerPlots, Day 4]

Application of models to compare two characteristics. For the second MAA of the first model development sequence, groups of students adapted their models of sampling and inference to compare two characteristics in a population. The goal of the MAA was to apply the model developed in the MEA to compare two populations and use simulated data to support or oppose claims regarding the populations from which the data was drawn. Students were told that the Adidas global share of the sneaker market is 20%. Nike is interested to know if this difference of about 15% between Nike and Adidas sneaker purchases was true for sneaker purchases by the
local university students. The students were told that in addition to the clear beads in the bin representing Nike sneaker purchases, the blue beads were Adidas sneaker purchases. Nike asked the groups to determine the likely range for the difference in percentage between Nike and Adidas sneaker purchases by local university students.

I classified the models and sampling and inference developed by the groups of students in two ways, by the groups’ methods of sampling and then by their methods of inference. Groups collected samples using two different methods. Focus groups B and C adapted their methods of sampling to compare the two characteristics by only counting one characteristic in each of their samples. These groups collected separate samples to count the Nike and then Adidas sneaker purchases. The remaining groups, including focus groups A and D, counted both characteristics in each sample by using one sample of sneaker purchases to count the numbers of Nike and Adidas sneaker purchases.

Groups then used three methods to compare the percentage of Nike and Adidas sneaker purchases, regardless of the method of sampling. In the first method of inference, comparing of means (n=12, including focus groups B, C, and D), groups found the mean number of Nike and Adidas sneaker purchases and used these means to draw conclusions about the difference in sneaker purchases at the university. In the second method of inference, comparing of likely ranges (n=5), groups constructed ranges of likely values for each Nike and Adidas sneaker purchase and compared these ranges to draw conclusions. For the third method of inference, comparing of differences in each sample, (n=2, focus group A), groups found the differences in Nike and Adidas sneaker purchases in each sample and used these differences to draw their conclusions.
Methods of sampling two characteristics. Two of the groups (focus groups B and C), collected 40 total samples of 20 sneaker purchases. In the first 20 samples, the groups counted the Nike purchases in each sample. In the remaining 20 samples, the groups counted the Adidas purchases in each sample. The remaining groups (n=17) counted the numbers of Nike and Adidas sneaker purchases from each collected sample. Focus group C presented their method of sampling to their class, which led to a discussion of how these two methods of sampling were different.

Zayn: So what they basically did was sample 20 [sneaker purchases], which was the white beads, then they only counted just the white beads, with that sample of 20 [sneaker purchases]. Then they did a whole another sample, and just counted all of the blue beads. But, is that sort of like, biased in a way? Like, because they’re just neglecting the other beads.

Brenda: Yeah, what does anybody think about that? So I think Zayn has maybe a point that we need to talk about here a little bit, because what Zayn is saying is these other groups, you scooped in [the bin of beads], and in that same trial you counted the blue and the white beads. Am I hearing you correctly?

Zayn: Yeah.

Brenda: And then you’re saying is what they’re doing (focus group C) somewhat different they’re taking, they essentially used 40, but they used 20 for the blue and 20 different ones for the white.

Zayn: Because they’re neglecting the results that they got in the white, they’re neglecting the blues and the oranges that they got. [Class discussion, Nike vs. Adidas, Day 5]
Zayn demonstrated an understanding of the dependence of the number of Nike and Adidas sneaker purchases in groups of 20 sneaker purchases. After counting the number of Nike sneaker purchases in one sample, that will affect the possible number of Adidas sneakers in the sample. As a class, the students decided with no further discussion that while these two methods differ, they would result in the drawing of the same conclusions. Brenda, Allen, and I did not anticipate that this model and discussion would occur.

*Methods of inference: Comparing of means.* The models constructed in this category used the mean numbers of Nike and Adidas purchases out of 20 sneaker purchases and used two methods to draw conclusions. In the first method, groups (n=9, including focus groups B and C) found the difference between the mean numbers of Nike and Adidas purchases out of 20 sneaker purchases and used this value as their predicted difference between the percentage of all Nike and Adidas sneaker purchases for students at the university. Focus group B used their data to assert that:

After the simulations our data demonstrated that 10/20 students at [local university] wore Nike shoes. Our other half of data stated that 9.35/20 students at [local university] wore Adidas shoes. The percent difference between these two numbers is 2.5%. [Written classwork, Nike vs. Adidas Market Share, Day 5]

Focus group B’s conclusion did not return to the initial claim by the Nike marketing director of the 15% difference in the market share of the two brands of sneakers. Rather than comparing their results to 15%, they stated that their “data demonstrated” that the difference was 2.5%. They misinterpreted what the beads represented. The beads represented sneaker purchases, and not the brand of sneakers worn by one student. One student could have made numerous sneaker purchases and be represented by more than one bead.
Focus group C was the only group in this category of model that represented their data in dotplots. Although the group constructed two dotplots, the plots were not used in presenting their conclusions. Brenda observed that focus group C was collecting samples and only counting the numbers of one brand of sneaker [Instructor interview, Day 5] and asked the group how they were approaching the MAA.

Brenda: So can I ask you for a second what, what you’re recording?

Megan: We’re just scooping up the 20 beads, the beads 20 times, and we’re counting the number of Adidas.

Brenda: Uh huh.

Megan: And then we’re going to compare the mean of this one to the mean of the Nike one, which was 10, and so if it’s like 15% between, so it’s 3 right? The difference?

Brenda: That’s what the, right, so the 15% difference.

Megan: So if it’s a 15% different then it will be 7. So if it holds true, then this mean will be 7. [Group discussion, Nike vs. Adidas Market Share, Day 5]

Megan noted what mean number or Nike sneaker purchases out of 20 sneaker purchases she would expect to observe if there was a difference of 15% in the two brands of sneaker purchases. Focus group C found the mean number of the Nike sneaker purchases to be 10/20, mean number of the Adidas sneaker purchases to be 7/20, and the difference to be 3/20. This led the group to assert that the difference between Nike and Adidas sneaker purchases among local university students was 15% and the Nike marketing director’s “statement is true” [Written class work, Nike vs. Adidas Market Share, Day 5].

In the second method in this category of model the groups (n=3, including focus group D) also found the difference between the mean numbers of Nike and Adidas sneaker purchases out
of 20 sneaker purchases, but then used this mean difference to construct an interval for the likely
difference in the percentage of Nike and Adidas sneaker purchases. Focus group D collected 10
samples of 20 sneakers purchases and created a table to count the numbers of Nike, Adidas, and
other brands’ sneaker purchases (Figure 26).

![Table showing counts of Nike, Adidas, and other brands' sneaker purchases](image)

**Figure 26.** Focus group D’s samples of Nike and Adidas sneaker purchases. [Written classwork, Nike vs. Adidas Market Share, Day 5]

Focus group D then averaged the numbers of each brands’ purchases and found that 42% of their
samples were Nikes, 30% Adidas, and 28% sneakers from some other brand. The group
subtracted the means of Nike and Adidas purchases and determined that there was a 12%
difference in Nike and Adidas sneaker purchases in their samples. The group then discussed how
to interpret this value.

Randy: The difference is 12% difference, right? And they thought 15% difference. They
wanted to know if it was accurate.

Ted: So, that’s really close, but it’s not exactly it.

Randy: Right.
Ted: So to say if it’s accurate or not. I remember her saying in one of the exercises where it wasn’t exactly, where she said you could argue that in the real world, and this is kind of like the real world. [Group discussion, Nike vs. Adidas Market Share, Day 5]

Ted is thinking about the variability in groups of 20 sneaker purchases. The group calculated a mean difference that was close to 15%, which led Ted to question if 12% and 15% are close enough to consider the Nike marketing director’s claim accurate. As this discussion wrapped up, Brenda addressed the class to direct them to consider the likely range of differences in Nike and Adidas sneaker purchases.

Ted: When you say the range of the difference in percentage, do you only mean just one percentage? Because a range is like multiple things and we’re only getting one percentage.

Brenda: So think about the likely range.

Susan: Ok, so you could say possibly that 15 could be in the range, but it’s kind of a stretch, ‘cause we got 12%.

Ted: Which is close to 15, but …

Susan: I would say maybe 15 is on the end of the range.

Brenda: On the high end of the range?

Ted: The range could be 12 to 15. Maybe the beginning and end of the range. [Group discussion, Nike vs. Adidas Market Share, Day 5]

The group came up with a range, 12% to 15%, that was not symmetric around the difference of their samples means. This led the group to discuss what the dotplot of the sampling distribution would look like if this were the range of likely values. Although they did not create a dotplot
with their 10 samples, they discussed that if many samples were collected, the plot would have a peak in the middle of the predicted range at 13.5% difference.

Susan: So our range is going to be 12 to 15%.

Ted: But that means we’re assuming that the point is 13.5 (motioning with his hands the top of a hill).

Susan: If you do this on a bigger scale. [Group discussion, Nike vs. Adidas Market Share, Day 5]

Ted described how the likely range of values can be found on the dotplot of the sampling distribution, as was done in previous activities. He demonstrated an understanding that the dotplot used to draw this conclusion would be made up of the differences between Nike and Adidas purchases. They did not discuss why they considered their difference of the sample means, 12%, and the Nike marketing director’s claim, 15%, to be equidistant from the peak of this distribution at 13.5%.

Coming into this MAA, focus group D constructed a model of inference to determine the likely range of outcomes by constructing a dotplot of the data, finding the center of the distribution, and choosing a range of values above and below this center. In this MAA they did not apply that reasoning to construct a likely range of values because there was no dotplot to examine. Having two sets of data (Nike and Adidas), rather one than, may have led them not to use this reasoning to evaluate their answer. When constructing a likely range from 12% to 15%, Ted realized that this meant that the center of the distribution was in the middle of those values, 13.5%, but this conflicted with their calculated average of 12%.

The remaining two groups in this category of model constructed their likely range of values for the difference in Nike and Adidas purchases in a different way than focus group D.
One of the non-focus groups collected five samples of 20 sneaker purchases and counted the number of Nike and Adidas purchases in each sample. They calculated the percentage of Nike and Adidas purchases in each sample, found the difference between these two percentages for each sample, and calculated the average of these differences. The group found the average of the differences to be 36% and concluded that the likely range for the difference between Nike and Adidas sneaker purchases at the local university would be 32% to 40%. Unlike focus group D, this group chose to construct an interval with the average of the sample differences in the middle of the likely range. To justify their likely range of values, the non-focus group sketched a bell shaped curve (Figure 27) with 36% in the center of the curve and lines drawn to the right and left of 36% marked for the end of the likely range, at 32% and 40%.

*Figure 27. Non-focus group’s sketch of the anticipated spread of samples for the difference between Nike and Adidas sneaker purchases. [Written classwork, Nike vs. Adidas Market Share]*

As discussed previously with focus group D, this group appears to be drawing on the methods used to construct likely ranges of outcomes in earlier activities with empirical sampling distributions. The non-focus group did not use their five samples to construct an empirical sampling distribution and did not give justification for why the values of 32% and 40% appear in the sketched locations on their curve (Figure 27).
Methods of inference: Comparing of likely ranges. Like the previous method of inference, groups in this category (n=5) constructed likely ranges of values. Groups constructed two dotplots, one for the numbers of Nike sneaker purchases in each sample of 20 sneaker purchases, and then a second for the number of Adidas sneaker purchases in each sample of 20 sneaker purchases. From these dotplots the groups constructed likely ranges of sneaker purchases for each brand and then compared these likely ranges to draw conclusions about the difference in the percentage of the two brands of sneaker purchases. Each group constructing this category of model followed a similar procedure. One non-focus group stated that:

We first used the bin of beads, and designated the clear as Nike and the blue as Adidas. We then did 20 simulated trials. After performing the trials, we found that Adidas was likely to be worn 3-8 times out of 20 students, and Nike was worn 6-12 times out of 20 students. [Written classwork, Nike vs. Adidas Market Share, Day 5]

This non-focus group counted the number of Nikes and Adidas sneaker purchases in each sample and constructed dotplots for each brand of sneaker (Figures 28 and 29). To construct the likely ranges of sneaker purchases for each brand, the group excluded the most extreme values that only occurred once and kept the remaining numbers of Nike and Adidas sneaker purchases to form the likely ranges. For the dotplot of the Nike data (Figure 28), the group included the outcomes of six and seven Nike sneaker purchases out of 20 purchases in their likely range although they only occurred once in the observed samples. For the Adidas dotplot (Figure 29), the samples with lowest and highest numbers of Adidas sneaker purchases in samples of 20 sneaker purchases were only observed once and not included in the likely range of values. A sample with three Adidas purchases of the 20 was also observed once, but was still included in the likely range of values.
Figure 28. Non-focus group’s dotplot for the number of Nikes in 20 samples of 20 sneaker purchases. [Written classwork, Nike vs. Adidas Market Share, Day 5]

Figure 29. Non-focus group’s dotplot for the number of Adidas in 20 samples of 20 sneaker purchases. [Written classwork, Nike vs. Adidas Market Share, Day 5]

Four of the five groups comparing likely ranges (including the non-focus group discussed previously) used these two likely ranges of values to construct a likely range for the difference in Nike and Adidas sneaker purchases. The remaining group did not use their likely ranges for Nike and Adidas sneaker purchases to draw any conclusion about the difference in the percentage of purchases. Returning to the non-focus group discussed above, the group determined an interval for the differences by comparing their two intervals of likely ranges. The group stated that:
We then subtracted the high to low ratios respectively which gave us our difference range of 3-4. We then multiplied these results by 5 to get our percent range of 15-20%. [Written classwork, Nike vs. Adidas Market Share, Day 5]

Each group that determined a likely range for the differences in Nike and Adidas purchases followed this same procedure of subtracting the upper and lower endpoints of the Adidas range of values from the upper and lower endpoints of the Nike range of values, respectively.

This method of comparing likely ranges of values is troublesome since it only compares the high end of likely values for Nike purchases with the high end of likely values for Adidas sneaker purchases, and the low end of likely values for Nike purchases with the low end of likely values for Adidas sneaker purchases. If all outcomes in the ranges are likely, it would be likely that a value from the low end of one range could occur, while a value from the high end of the other range occurs. Returning to the non-focus group’s data (Figures 28 and 29), the group determined that the likely range for Nike purchases was six to 12 for 20 sneaker purchases and three to eight for Adidas purchases in 20 purchases. This implies that it would be likely to sample 20 sneaker purchases and find that 12 were Nikes and three were Adidas. The difference between these purchases is nine sneakers or 45%. This percentage is outside of the group’s likely range for the difference in Nike and Adidas purchases, 15%-20%. This non-focus group’s range of 15%-20% is a prediction for the difference in the center of the two distributions of data rather than the ranges of likely values.

Methods of inference: comparing of differences in each sample. The remaining groups (n=2, including focus group A) collected samples of 20 sneaker purchases by counting the number of Nike and Adidas sneaker purchases in each sample and finding the difference between these values. Unlike previously discussed models, groups in this category constructed a dotplot
of the differences in purchases and determined a likely range for the difference in Nike and Adidas sneaker purchases from this dotplot.

Focus group A collected 10 samples of 20 sneaker purchases and constructed a table with the proportion and percentage of Adidas sneakers and the number and percentage more Nike purchases than Adidas (Figure 30). The focus group recorded the number of Adidas sneaker purchases for each sample in the table and subtracted these numbers from the Nike sneaker purchases to determine how many more Nikes were in each sample than Adidas. This difference between the numbers of sneakers from each brand as also recorded as a percentage for each sample.

<table>
<thead>
<tr>
<th>Trial</th>
<th>Difference</th>
<th># of Adidas</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0%, None</td>
<td>7/20 = 35%</td>
</tr>
<tr>
<td>2</td>
<td>30%, 6 more Nike</td>
<td>2/10 = 10%</td>
</tr>
<tr>
<td>3</td>
<td>10%, 2 more Nike</td>
<td>7/20 = 35%</td>
</tr>
<tr>
<td>4</td>
<td>15%, 3 more Nike</td>
<td>6/20 = 30%</td>
</tr>
<tr>
<td>5</td>
<td>10%, 2 more Nike</td>
<td>5/20 = 25%</td>
</tr>
<tr>
<td>6</td>
<td>35%, 7 more Nike</td>
<td>4/20 = 20%</td>
</tr>
<tr>
<td>7</td>
<td>10%, 2 more Nike</td>
<td>7/20 = 35%</td>
</tr>
<tr>
<td>8</td>
<td>10%, 2 more Nike</td>
<td>5/20 = 25%</td>
</tr>
<tr>
<td>9</td>
<td>20%, 4 more Nike</td>
<td>7/20 = 35%</td>
</tr>
<tr>
<td>10</td>
<td>30%, 6 more Nike</td>
<td>6/20 = 30%</td>
</tr>
</tbody>
</table>

*Figure 30. Focus group A’s table of how many more Nikes were in their samples compared to Adidas sneakers. [Written classwork, Nike vs. Adidas Market Share, Day 5]*

Focus group A determined the average of the percentage difference between Nike and Adidas purchases for each sample, 17%, and initially constructed a likely range for the differences of 15-20%. The group chose values slightly above and below the average difference of 17% to determine this likely range and did not give reasoning for why those values were chosen.
This method was in line with the second subcategory of model, where the groups found the average and used this average to construct a range of values. Focus group A then constructed a dotplot with the values for how many more Nike purchases were in each sample compared to Adidas sneakers (Figure 31).

![Dotplot of Nike vs. Adidas Market Share](image)

*Figure 31. Focus group A’s dotplot of how many more Nikes were in their samples compared to Adidas sneakers. [Written classwork, Nike vs. Adidas Market Share, Day 5]*

From this dotplot, focus group A determined that the likely range of values was 2-6 more Nikes than Adidas. To construct this interval the group eliminated the highest and lowest value in their dotplot that only occurred once and kept the remaining values to form the likely range. From the dotplot (Figure 31) the group determined that their initial likely range would only include the outcomes for three and for more Nikes. This led the group to construct a likely range of values as they had done in previous activities, by including most of the observed values in the dotplot, while excluding the most extreme values that only were observed once.

**Summary of focus groups’ models of sampling and inference.** Next is a brief summary of the developing models of sampling and inference constructed by each focus group over the first model development sequence.
Focus group A. This group began the unit by simulating data for the MEA with eight coins. All of the outcomes that they observed when simulating ten sets of Ophelia’s predictions were viewed as likely outcomes. The group asserted that if Ophelia correctly predicted more games than the values in their range of likely outcomes (two to five correct predictions out of 8), then the octopus may not have been guessing the winners of each game outcomes. When each of the three group members simulated 10 sets of outcomes for Ophelia in the MXA, group members were surprised that each of their dotplots were different from one another and did not peak in the middle. These thirty sets of predictions were combined to create a new range of likely values that again contained all observed values. The group noted that this interval was wider than the interval created in the MEA. The group members discussed ideas of variation when simulating different numbers of trials. When simulating fewer trials, the group members asserted that it was less likely to observe more extreme outcomes. The group simulated 1000 trials and no longer viewed all observed values as likely. A range of likely values was constructed according to outcomes that occurred the most, centered on the peak of the sampling distribution.

In the first Nike sneakers MAA, focus group A examined multiple aspects of their sampling distribution to draw conclusions. They first claimed that since all of their initial 10 samples were greater than or equal to the claimed value of $7/20$ Nike to total sneaker purchases, there was evidence that $7/20$ was too low for the proportion of Nike sneaker purchases at the university. The group later examined 105 samples and determined the likely range of Nike sneaker purchases by excluding a few extreme values, but decided to use the center of the dotplot to draw their conclusions. The claimed value of $7/20$ Nike to total sneaker purchases was below the center of $9/20$ Nike to total sneaker purchases, and therefore too low. To compare Nike and Adidas sneaker purchases in the second MAA, the group counted the number of Nikes and
Adidas sneaker purchases in each sample, and then calculated the difference in purchases. These difference values were aggregated in a dotplot. The group excluded the extreme values of differences that only occurred once and used the remaining values to construct a likely range for the difference in the two brands’ sneaker purchases.

*Focus group B.* This group began the MEA by simulating outcomes for Ophelia’s predictions using the eight coins. They determined a likely range of outcomes consisting of all outcomes that they observed in their simulation. When individually simulating data with *TinkerPlots* in the first MXA, the students noted that with 10 samples, each of their dotplots’ observed outcomes looked very different. Because of these differences, they did not know what to conclude about the likely range of values. When simulating 50 trials their dotplots were much more similar and for the first time constructed ranges of likely values from outcomes that occurred the most often in their dotplots. After simulating 1000 trials, the group members asserted that their dotplots looked identical.

For the first MAA, focus group B collected samples and determined a likely range of values containing the outcomes that occurred most in their samples. This range was compared to the claimed value and since the claimed value was below their range, the group asserted that the claim was too low. Continuing in the Nike context with the MXA, the group emphasized that they needed to collect a large number of samples so their dotplot would look like a mountain, implying the need for a consistent bell curve shaped distribution. In the second MAA, the group took separate samples when counting the numbers of Nike and Adidas sneaker purchases. The group compared the mean numbers of Nike and Adidas sneakers for each sample, found the mean different to be 2.5%, but did not relate the difference in means to the claimed difference of 15%.
Focus group C. The group began the model development sequence hesitant to use the coins as manipulatives, but eventually used them to simulate Ophelia’s predictions in the MEA. The group determined that the simulated data were inconsistent, since more than one outcome for the number of correct predictions was observed. This led the group to assert that they could not draw a conclusion. When simulating only 10 trials in the first MXA, the group members’ TinkerPlots environments did not generate bell shaped dotplots as they has expected. When 50 and then 1000 trials were simulated, group members found that the dotplots looked more like bell curves the more trials were simulated. Going beyond 1000 trials, they asserted that at some point the dotplot looked so much like a bell curve that collecting more samples had no effect on its shape.

For the first MAA, the group found a likely range of values for Nike sneaker purchases that consisted of most values around the center of their distribution. The Nike marketing director’s claimed value was below this interval, which led to the group to assert that the claimed value was too low. When using TinkerPlots in the second MXA, the group continued to assert the need for a large number of samples (1000) in order to produce a bell curve shaped distribution. To compare the Nike and Adidas sneaker purchases in the second MAA, the group collected separate samples to count the numbers of each brand of sneaker. They compared these groups of samples by calculating the mean numbers of each brand and determined the difference in the means. This one value of the difference, rather than an interval as they used in previous activities, was used to compare to the Nike marketing director’s claimed value.

Focus group D. The group began the model development sequence by using the coins to simulate Ophelia’s predictions in the MEA and constructed a likely range of values that consisted of all values observed in their trials. When using TinkerPlots to simulate data in the first MXA,
the group no longer used all observations in their likely ranges and instead used the outcomes that occurred the most often. One group member examined the proportions of outcomes in the likely ranges and based the width of his range on capturing what he deemed a suitable proportion (0.74) of the total number of simulated samples.

For the first MAA, the group collected 10 samples and discussed the distribution of the values on a dotplot, but decided to calculate the mean number of Nikes in their samples to use as evidence. The mean value was asserted to be their predicted value for all sneaker sales at the university and was not compared to the Nike marketing director’s claim. When using TinkerPlots to simulate the Nike sneaker data in the second MXA, the group noted the need to collect a large number (1500) of samples. One group member asserted that when they each simulated a large number of samples their dotplots would look the same, but that was not true for smaller numbers of samples. Large numbers of samples meant consistency among their predictions. To compare Nike to Adidas sneaker purchases in the second MAA the focus group counted both brands’ sneaker purchases in the samples and determined the difference in purchases for each sample. From these differences, the group calculated the mean of the differences, which was close to the claimed difference. The group claimed that the likely difference between the sales of the two brands was between their calculated mean difference and the claimed mean difference.

Models of Resampling and Inference

In the second model development sequence, groups extended their previous models for sampling and inference by constructing resampling methods. This sequence no longer had an available population from which to draw many samples. The groups of students had only one sample from the population that could be used to make inferential claims (see Figure 3). The model development sequence elicited and developed the method of bootstrapping.
uses one sample of data from the population to construct a bootstrap empirical sampling
distribution, similar to the sampling distribution used by groups of students in the first model
development sequence. From the initial sample, which is assumed to represent an approximation
of the population's distribution, bootstrap samples are constructed by choosing \( n \) elements, \textit{with}
replacement, creating many equal sized samples. Sampling with replacement is used under the
assumption that the data values in the sample represent the distribution of all data values in the
population and the proportion of each data value in the sample represents the proportion of that
data value in the population. A statistic from each of these bootstrap samples is then calculated
and aggregated to form an empirical bootstrap sampling distribution (see Figure 2).

**Initial models of resampling.** The MEA of the second model development sequence was
designed to elicit the idea of bootstrapping by resampling with replacement. The goal of the
MEA was to elicit a model that was used to resample from a sample, with replacement, in order
to construct a bootstrap sample of the same size as the original sample and use the distribution of
these bootstrap samples to determine which are most likely to occur. Students were asked to help
the manager of a grocery store predict the percentage of peanuts in a certain brand of mixed nuts
(Appendix C). The students were given a sample of mixed nuts in the form of seven craft sticks
marked with a “P” for peanut and 18 not marked to represent other kinds of nuts. The manager
planned to buy a large shipment of mixed nuts, but thought that her customers preferred brands
with fewer peanuts. From this one sample of mixed nuts, the manager asked the students to
determine a likely range for the percentage of peanuts in the overall brand of mixed nuts. This
activity was in contrast to activities in the first model development sequence since the students
did not have the option to take additional samples of mixed nuts from the population of mixed
nuts in order to draw their conclusions.
Four categories of models for resampling and inference were constructed by the groups of students (Table 1). The first category of model (n=14 groups, including focus groups A, B, and C) treated the sample of 25 nuts in a manner similar to a population and collected resamples of the same size (either five or 10 nuts), without replacement, from the 25 nuts. The second category of model (n=2 groups) also sampled without replacement, but varied the size of each resample. The third category of model (n=3, including focus group D) discussed or collected resamples from the sample of mixed nuts by choosing with replacement. The fourth category of model (n=1) created a spinner, similar to what was used in first MXA in the first model development sequence, with areas of a pie chart proportional to the number of peanuts and other nuts in the group’s sample. The group then simulated resamples of 15 nuts, one spin at a time.
Table 6

Models of Resampling.

<table>
<thead>
<tr>
<th>Model Category (Groups)</th>
<th>Characteristics of Model Category</th>
<th>Model Subcategory Characteristics (Groups)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resampled without replacement: Resamples of equal sizes (n=14, A, B, C)</td>
<td>Collected 10-20 resamples without replacement, each containing 5 or 10 nuts.</td>
<td>Range’s end points were not equidistance from the mean and were rounded to the tens of a percentage. (n=3, A)</td>
</tr>
<tr>
<td></td>
<td>Constructed a likely range for the peanuts in the new brand.</td>
<td>Range’s end points were equidistance from the mean. (n=2, C)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Range contained all values observed in their resamples. (n=5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Range contained most values in the resamples. Extreme values that occurred least often were excluded. (n=4, A, B)</td>
</tr>
<tr>
<td>Resampled without replacement: Resamples of varying sizes (n=2)</td>
<td>Collected 10 resamples, varying in size from 9 to 18.</td>
<td>Range’s end points were roughly equidistance from the mean. (n=1)</td>
</tr>
<tr>
<td></td>
<td>Constructed a likely range for the peanuts in the new brand.</td>
<td>Range contained most values in the resamples. Extreme values that occurred least often were excluded. (n=1)</td>
</tr>
<tr>
<td>Resampled with replacement (n=3, D)</td>
<td>Collected (or discussed collecting) resamples of equal sizes, with replacement.</td>
<td>Collected two resamples of 20 nuts, but did not use the data to draw a conclusion. Resampling was viewed as inefficient and did not yield expected results. (n=1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Discussed, but the group did not collect 25 samples of 10 nuts. Resampling with or without replacement were viewed as yielding the same results. (n=1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Collected one resample of 14 nuts before running out of time. The group discussed that resampling preserved the makeup of the sample when selecting each nut. (n=1)</td>
</tr>
<tr>
<td>Resampled with spinner (n=1)</td>
<td>Constructed a spinner, with two regions equivalent to the proportion of peanuts and other nuts, to simulate samples. Collected 4 resamples of 15 nuts, but did not use the data to draw a conclusion. (n=1)</td>
<td></td>
</tr>
</tbody>
</table>
Resampled without replacement: Resamples of equal size. Each group that constructed a model in this category collected resamples from the sample of 25 nuts, without replacement, with equal sized resamples of either five or 10 nuts. The variation between these models came from how the groups of students used this data to draw conclusions. Four subcategories of models were used to draw conclusions:

1) Groups found the mean number (or percentage) of peanuts in their resamples and constructed an interval around this mean with end points that were *not* equidistant from the mean (focus group A).

2) Groups found the mean number (or percentage) of peanuts in their resamples and constructed an interval around this mean with end points that were equidistant from the mean (focus group C).

3) Groups constructed an interval that contained *all* of the values for the number of peanuts in their resamples.

4) Groups constructed an interval that contained *most* of the values for the number of peanuts in their resamples (focus groups A and B).

Some groups’ models, such focus group A, developed over the course of the activity and constructed models from more than one subcategory.

Focus group A collected 10 resamples of 10 nuts, which they displaying initially in a table and then later with a dotplot (Figure 32). The group calculated the mean (32%) for the percentages of peanuts in their resamples and concluded from the mean that there was a likely range of 20-40% peanuts. During class discussion, Allen asked the group how they decided on this range of values.

Allen: I noticed that there were some numbers [of peanuts in a sample] that you didn’t
include [in the likely range for the percentage of peanuts]. You didn’t have a
dotplot [focus group A later constructed a dotplot], but there were some numbers
that you didn’t include.

Cara: Well …

Allen: ’Cause you just felt they were outside?

Cara: Well, we based our range on our average.

Allen: Oh, okay, so they took the average and just kind of went to the whole tens on
either side.

Cara: Right. [Class discussion, Mixed Nuts, Day 6]

This focus group initially used the sample mean as the center of their predicted likely range, and
chose values that were roughly symmetric above and below this mean that were rounded to the
tens place.

<table>
<thead>
<tr>
<th>Trial #</th>
<th># of P (x10)</th>
<th>$%$</th>
<th>Avg $= 32%$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5/10</td>
<td>50</td>
<td>20 - 40%</td>
</tr>
<tr>
<td>2</td>
<td>3/10</td>
<td>30</td>
<td>20 - 40%</td>
</tr>
<tr>
<td>3</td>
<td>4/10</td>
<td>40</td>
<td>peanuts</td>
</tr>
<tr>
<td>4</td>
<td>4/10</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1/10</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>2/10</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>4/10</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>3/10</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>3/10</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>3/10</td>
<td>30</td>
<td></td>
</tr>
</tbody>
</table>

Figure 32. Focus group A’s table and dotplot for the counts of peanuts in 10 resamples of mixed
nuts, collected without replacement. [Written classwork, Mixed Nuts, Day 6]
Focus group C constructed the second subcategory of model in the first category (Resampled without replacement: Resamples of equal size). This group also used the mean of the resamples to draw their conclusions, but found the predicted interval in a different way than focus group A. Initially focus group C counted the sticks in the bag and found that seven were marked with a P and 18 were not marked. From this they determined that 28% of the nuts were peanuts. Megan asked the group, “Is that all we have to do for this?” The instructor Brenda then came over to talk with the group.

Brenda: Okay, so how is this group doing?

Megan: We just counted the number of peanuts, put it over 25, multiplied it by 4, got 28%

Brenda: Okay, and what are you going to do with that? How accurate do you think that your sample is? [Group discussion, Mixed Nuts, Day 6]

Brenda may have used the language of “accurate” to encourage the group to construct a likely range of values for the percentage of peanuts rather than predicting one singular value.

Megan: Probably not entirely [her emphasis] accurate because it’s only one sample, but I heard Walter [student from a non-focus group] saying he also got 28%, so if we went to other groups, there would be more samples for that. [Group discussion, Mixed Nuts, Day 6]

Megan expressed an idea of variability in sampling by showing uncertainty if this sample was a good representation of the percentage of peanuts in the entire population. By saying “not entirely accurate,” she may view this sample as likely to be similar to the population of nuts, with some smaller chance that it is not like the population. By suggesting that her group collect more samples from other groups, Megan was attempting to apply methods similar to those she had
used in the previous model development sequence where the population was available and many samples could be collected. In this case of an unavailable population, each group was given an identical sample, with seven of the 25 sticks marked with a P for peanut. Brenda continued to speak with the group.

Brenda: Is there a way, because you need to come up with a likely range, right? So is there another way that you can use that sample to help you? Because think about before when you were getting a likely range. You were taking samples, right? That was from the population, but is there something, some way that you could use your sample?

Megan: Like, grab five sticks and count how many peanuts there are?

Brenda: Well, I’ll let you talk about it. You don’t have to answer me. I’ll let you talk about that with your group (Brenda leaves).

Megan: So, would that work if I grab 5 sticks right now and count how many peanuts, and do that like 20 times or something or 5 times? … For the first one I got 0, so no peanuts. So I’ll put them back, right?

Karl: Yeah. [Group discussion, Mixed Nuts, Day 6]

Focus group C collected five resamples of five nuts, recorded the samples in a table (Figure 33), and calculated the mean percentage of peanuts in the five resamples.
Figure 33. Focus group C’s table for the count and percentage of peanuts in five resamples of mixed nuts, collected without replacement. [Written classwork, Mixed Nuts, Day 6]

To determine the range for the likely percentage of peanuts, the group first determined the difference between the mean percentage of peanuts in their resamples (32%) and the percentage of peanuts in the initial sample (28%) (Figure 34). The four percent difference in these two values was used to construct the range of 24% to 32% peanuts, which was the initial sample of 28% peanuts plus and minus the difference between the “observed” and “actual” values of four percent.

Figure 34. Focus group C’s likely range for the percentage of peanuts in the brand of mixed nuts.

[Written classwork, Mixed Nuts, Day 6]
Since the sample of mixed nuts contained 28% nuts, I interpreted that the group viewed this as the actual mean percentage of peanuts that they expected to collect in their resamples. When the group collected a mean percentage different from what they had expected, the group subtracted these values to determine how far away from the expected mean percentage was likely to be observed. This difference was four percent, which led them to conclude that they expect the actual amount to be four percent above or below the percentage that they expected to collect, 28%. The group collected two additional samples, constructed a dotplot, and calculated the new mean percentage of peanuts from the seven resamples (Figure 35), but did not use this new data to change their claimed range for the percentage of peanuts in the brand of mixed nuts.

![Dotplot](image)

**Figure 35.** Focus group C’s dotplot for the counts of peanuts in 7 resamples of 5 mixed nuts, collected without replacement. [Written classwork, Mixed Nuts, Day 6]

Five groups constructed models in the third subcategory of Resampled without replacement: Resamples of equal size and created intervals for the likely range of peanuts that contained all observed values in their resamples. One non-focus groups collected 14 resamples of 10 nuts and displayed this data in a dotplot (Figure 36). The group drew brackets around their predicted likely range of peanuts, which contained the number of peanuts in all 14 resamples.
One of these five non-focus groups, while predicting a likely range of peanuts that contained all of their outcomes, collected results that differed slightly from the other four groups. Four of the five non-focus groups had outcomes in their predicted ranges that only occurred once, but in the remaining group’s samples, each outcome that occurred, occurred more than once (Figure 37).

Since none of the group’s outcomes occurred only once, I could not specify if the group used their model of resampling and inference to construct a likely range based on all occurring
outcomes, like the four other groups, or if the likely ranges contained outcomes that occur most often or more than once, as in the next subcategory of models.

In the fourth subcategory, groups of students constructed likely ranges of peanuts that contained most of the outcomes from their resamples. Focus group B collected 17 samples of 10 nuts and displayed that information in both a table and dotplot (Figure 38). The group wrote on the worksheet “5-10 range”.

![Figure 38](image)

*Figure 38. Focus group C’s table and dotplot for the counts of peanuts in 17 resamples of 10 mixed, collected without replacement. [Written classwork, Mixed Nuts, Day 6]*

I interpreted that they were predicting between five and 10 peanuts in a sample of 25 nuts, for a range of 20% to 40%. From the group’s dotplot (Figure 38) all values except for one fall in the outcomes with two, three, or four peanuts. These values correspond to 20% to 40% peanuts in resamples of 10 nuts. I surmised that they excluded the outcome of one peanut from their interval since the other three outcomes occurred with a greater frequency.

Focus group A’s model of resampling continued to develop over the course of the activity. After developing their initial model (from the first subcategory, resamples of 10 nuts without replacement, using the mean numbers of peanuts in the resamples) focus group A constructed a
The dotplot showed that most of the resamples contained three or four peanuts and outcomes of one, two, and five peanuts only occurred once. The group used the dotplot to claim a different range for the likely percentage of peanuts than they had previously determined (initially 20%-40%). Similarly to focus group B, focus group A determined that since the outcomes of three and four occurred the most in their resamples, this provided evidence that 30% to 40% of the mixed nuts were likely to be peanuts. The group noted “*No replacement” on the worksheet to distinguish these results from a later revision to this model of resampling, that included sampling with replacement, after group discussion of a follow-up activity.

*Resampled without replacement: samples of unequal size.* One non-focus group, from Allen’s class with focus group B, decided to take samples of varying sizes in order to have more variation in the samples.

Xavier: We decided that since if you just pull out the same amount each time you would get the same answer obviously.

Allen: If we took 25, if we took all of them.

Xavier: Yeah, so we decided to do random trials with different amounts, so that we could get a more varied group. So, that’s just basically what we did. So then we got the percent of the amount of peanuts.

By varying the sizes of the samples, the group thought that they could collect a broader range of samples than those with a constant sample size. The second non-focus group using this method, from Brenda’s class with focus group C, asserted that varying the resample size was a more realistic scenario. When buying bulk food, such as nuts, you don’t purchase a certain number of nuts, but rather a scoop taken from a large bulk bin. During the presentation of this group’s
model Yon stated, “we’re doing ten trials and we’re grabbing a group. When you grab actual bulk food you don’t know the certain size you’re taking, you can’t just grab one [size]” [Class Presentation, Mixed Nuts Activity, Day 6].

To determine the percentage of peanuts in this brand of mixed nuts Yon’s group collected 10 resamples, which varied in size from nine to 18 nuts, and organized them into a table (Figure 39). The group found the percentage of peanuts in each resample, averaged these percentages together, and found the mean percentage of peanuts in their resamples to be 29.4%. The group used this mean percentage to predict that the likely range of peanuts in the brand was 15% to 44%, which is approximately 15% above and below the mean percentage of peanuts in their resamples. The group did not discuss why they chose this range of values.

![Table with data](image)

*Figure 39. Yon’s non-focus group’s table for the number of peanuts and other nuts, collected without replacement, with varying resample sizes. [Written classwork, Mixed Nuts, Day 6]*
The first non-focus group discussed in this category (from Allen’s class) displayed their data in both a table and a dotplot. The dotplot that William’s non-focus group presented to the class varied from all other groups’ dotplots. Each other group used the number of peanuts in their resamples as the units. This non-focus group used the percentage of peanuts in their resamples instead.

Allen: …you had to do yours [dotplot] differently. How come?

Xavier: Because we had to show the percent of peanuts instead of the number of peanuts out of how many trials

Allen: How come it didn’t make sense for yours [dotplot] to put the number of peanuts?

Xavier: Because we did different amounts in trials [Class discussion, Mixed Nuts, Day 6].

The group demonstrated an understanding of the number of peanuts in resamples of varying size not being equivalent to percentages of peanuts on those resamples.
The non-focus group used this dotplot (Figure 40) to determine that the likely range of peanuts for this brand was 25% to 35%. This range included all but two of their outcomes from their set of resamples. The excluded outcomes each only observed once and the smallest and highest percentages of peanuts observed.

*Resampled with replacement.* Three groups discussed ideas of resampling with replacement, including focus group D. One non-focus group from the Brenda’s class with focus group C began by choosing sticks one at a time from the bag with replacement. They chose not to pursue the method because they received results that were unexpected.

Walter: First we were just picking out individuals, and, the majority of those were not peanuts.

Brenda: Ok, so tell me exactly how you were picking those individuals.

Walter: We’d pick one out, put it back in, after we get our result, pick one out, put it
back in. Every time, we were just picking up the one that's not peanut, so we just thought that it wasn’t going to be as efficient.

Brenda: So you were planning on doing 20 of those, right? [20 resamples of 20 nuts]

Walter: Yup.

Brenda: Did you finish 20?

Walter: No, we stopped it.

This non-focus group collected two resamples of 20 nuts, by selecting each nut one at a time with replacement, and calculated the proportion of peanuts in in each of the two samples (Figure 41). The grouped marked a “P” in the table when they chose a peanut and an “O” when another nut was chosen. The table is scribbled through to denote that the group did not continue using this method to draw their conclusion.

Figure 41. Non-focus group’s data for counting peanuts and other nuts in two resamples of 20 mixed nuts, collected with replacement. [Written classwork, Mixed Nuts, Day 6]
The group asserted that the results were unexpected since the proportions of peanuts in each sample (2/20 and 5/20) were lower than they had expected. If the proportion of peanuts in this new brand of nuts was 28% (as it was in the group’s initial sample) it would be very unlikely to collect a resample of 20 nuts, with replacement, that only has two peanuts. Because the group collected an unlikely sample, they thought that their approach to the task was incorrect. The non-focus group also noted that the process was inefficient. When Brenda spoke to their group earlier in the class, the group had planned to collect 20 resamples of 20 nuts. Resampling with replacement is more time consuming than resampling without replacement. The time consuming nature of their model of resampling was viewed by the group as inefficient and contributed to the group changing their approach to the activity.

A non-focus group, from Brenda’s class with focus group D, discussed ideas of resampling. William, whose group also constructed a model of resampling without replacement and varying resample sizes, offered a different method to approach the activity. During a class discussion he described the approach that he had wanted to use to go about the activity, but his group decided not to pursue.

Brenda: Talk about what you wanted to do. Your group wouldn’t let you do it

William: Yeah, basically, I kind of want to do the same thing, pick out 10, but do it by singles, just like, take out one, and then figure it out by then.

Brenda: Okay.

William: Instead of taking all 10 and counting, just pick out one and say if it’s peanut or other and do that, like 25 times. Putting it back in there each time and figuring out the likely range of it being other or peanut.

Brenda: So what about that feels better than, a better method for sampling than doing
what your group has done?

William: I wouldn’t say that it would be better; I would think that it would be the same. It was just a different method of doing it. [Class discussion, Mixed Nuts, Day 6]

William described a method with resampling with replacement and the collection of resamples that are the same size as the original sample. His reasoning of sampling with versus without replacement was still developing, but his group went in a different direction before he could pursue this reasoning.

Focus group D constructed a method of resampling similar to bootstrapping, but they had difficulty deciding how to proceed when they first approached the activity. Susan asked, “how do we want to figure out, like different ways to, because it’s not like you can flip the sticks or anything like that?” Group discussion, Mixed Nuts, Day 6] Susan was referring back to the MEA in the first model development sequence with Ophelia the Octopus. In that MEA there was not a physical population to draw from like the bin of beads in the Nike sneakers MAAs. In the Ophelia MEA, the coins could be flipped to simulate the outcomes. Susan did not think that the sticks could be used in a similar manner. Brenda approached the group and asked how they were going about the problem.

Susan: I would say from this sample, that a little over a quarter of the peanuts, of the nuts are peanuts, from our random sample.

Randy: But it’s only one sample.

Ted: But this is the only sample we have.

Brenda: So if you pick another random sample what’s going to happen?

Ted: It’s most likely going to change.

Susan: It’s going to change, but I feel like it will probably be still about the same.
Susan made an assertion about the percentage of peanuts in the population under the condition that she was basing her assertion on the random sample, which Randy stressed is only one random sample. Susan emphasized that although other samples of nuts will be different from their one sample, they will probably be “still about the same”. Susan demonstrated an understanding that the percentage of peanuts in the sample likely represents the percentage of peanuts in the population. This is a key aspect of inferential statistics. When making inferential claims you take for granted that the sample represents the population because as Ted stated, “this is the only sample we have” and as Susan asserted, “it will probably still be about the same” as other samples.

Brenda: So how could you do, what method could you use, to do something similar to that, so it felt like you were sampling from the population?

Susan: With the Popsicle sticks?

Brenda: With the Popsicle sticks. What could you do with the Popsicle sticks?

Susan: We could like mix them around and kind of like pick one out without looking and see like how often you pick out a peanut compared to like another nut.

From this excerpt Susan discussed taking resamples from the sample, but only one stick at a time. This is in contrast to what was done in previous MAAs with the Nike sneakers when all 20 sneaker purchases were sampled at once with the spatula.

The group decided to take a resample of 14 mixed nuts by drawing one at a time from the bag of mixed nuts, without replacement. They did not discuss why they resampled one at a time versus all 14 at once. The group focused on taking a resample size that was greater than one half
the size of the sample. They settled on 14 since it was more than 12.5 (half the size of the sample) and also an even number. Ted provided the reasoning, “if you did 12 and did two trials it would be 24 and it wouldn’t count for the whole sample.” Randy did not seem to follow or agree with this reasoning, but went along with taking resamples of 14 sticks. The group only took one sample with this method, which yielded five peanuts out of 14 nuts. Brenda then came back over to the group to discuss how the group had collected the sample of 14 nuts.

Brenda: So how are you choosing those?

Susan: He [Randy] randomly puts them together, and then I randomly without looking draw them out.

Brenda: Okay. So you’re drawing out one at a time?

Susan: Uh-uh.

Brenda: Okay, and you’re setting it on the table?

Ted: Yes.

Brenda: Then you’re going back in and you’re drawing another one?

Susan: Yes.

Brenda: Okay, so …

Ted: Ohh! Wait, you said that you’re putting them on the table. Was that like, do you think that we should put them back in the bag after we draw it out? Like for probability simulators? [Group discussion, Mixed Nuts, Day 6]

From this exchange Ted considered how the sampling would change if they resampled with replacement instead of resampling without replacement. By using the term “probability simulators” I think that he was approaching the idea that if you don’t replace the stick after choosing each one, the probabilities of choosing a peanut or another nut will change. Ted was
combining the idea of the representativeness of the sample to the population, that the group discussed earlier, with the idea of the probabilities staying constant for each choice of nut, the key concepts of bootstrapping. The group continued to discuss how this process of replacement was different than their initial approach without replacement. They asserted that when not using replacement, the nuts that they were drawing from the bag no longer represented the sample after some nuts were drawn and not returned.

Randy: What are we going to do now?

Susan: Now I’m going to hand you back the Popsicle stick and you’re going to mix it back in.

Randy: So you’re going to draw …

Susan: From the 25, not from, you know how before, like when I drewed [sic] and set it down, it went less and less and less and less and less?

Ted: So we’re going to do it again, we’re still going to draw 14, we’re just going to put them in.

Susan: So each time we’re drawing from 25, instead of a reduced …

Ted: ‘Cause as we would draw, in this one we would draw and there were 10, that meant that there were only 15 left in the bag, which doesn’t account for the sample, right? ‘Cause you’re reducing it. [Group discussion, Mixed Nuts, Day 6]

The group followed this procedure of resampling with replacement to collect a sample of 14 nuts, with five peanuts, and nine other nuts. This was the only sample that they had time to gather before the class was reconvened to discuss each group’s approach to determine the likely range of peanuts in this new brand. The group did not use this one sample to draw a conclusion
since time ran out for developing their model.

*Resampling with a spinner.* For the final category for models of resampling, one non-focus group constructed a spinner using a pencil and a hairpin (Figure 42).

*Figure 42. Spinner constructed by a non-focus group to resample the mixed nuts. [Written classwork, Mixed Nuts, Day 6]*

In the MXA Ophelia with *TinkerPlots*, a spinner was used to simulate Ophelia’s guesses when predicting the winner of one basketball game. Since there was an equally likely chance that Ophelia would guess correctly or incorrectly, the spinner appeared as a pie chart with half marked right and half marked wrong. The spinner was then spun eight times to determine Ophelia’s guesses for all eight basketball games. For this mixed nut activity, the non-focus group determined that since 28% of the sample was peanuts, they should construct a spinner with 28% of the pie chart marked peanuts, and the remaining are marked other nuts. The group then collected four resamples of 15 nuts (Figure 43).
The non-focus group discussed their approach when presenting it to the class:

Yvette: Basically what we did is we made a spinner. So we had 60 trials using the spinner, we did 15 four times [four samples of 15 nuts].

Allen: I don’t think that anyone else thought about a spinner. Can you tell us about, we haven’t seen a spinner since…can you tell us about how you came up with this spinner?

Yvette: We counted how many peanuts we had out of the 25 and 7… so that’s 28%, and then we constructed that spinner [Allen shows the spinner to the class Figure 42]

Allen: Hair pin, pencil, paper spinner, drawn approximately right, right? Slightly bigger than a quarter. [Class presentation, Mixed Nuts, Day 6]

From the four resamples (Figure 43) the group did not determine a likely range for the percentage of peanuts in the brand of mixed nuts. One of the resamples contained 13% peanuts, while the others contained 26.7%, 26.7%, and 33%. Yvette said that she was “thrown off” by the low value of 13% and wasn’t sure what conclusion to draw. The group viewed the sample with
13% peanuts as unlikely sample, which made them question the validity of the method that they used to simulate the resamples.

**Developing models of resampling.** After each group presented their models of resampling to the class a whole class discussion took place to compare and contrast the models. In Allen’s class with focus group A, no groups constructed models of resampling with replacement. During class discussion, Allen discussed possible issues with models of resampling without replacement, such as not being able to collect all possible samples. Allen’s class with focus group B, the model of resampling constructed with the spinner had structural similarities to resampling with replacement. This led Allen to discuss how using the spinner contrasted with resampling without replacement and how the spinner related to resampling with the craft sticks.

In each of Brenda’s classes (with focus groups C and D), groups pursued ideas of resampling with replacement. In Brenda’s class with focus group C, one group initially pursued resampling with replacement, but then segued to resampling without replacement before drawing a conclusion. Brenda chose to move to the follow-up activity (discussed next) and compare resampling with or without replacement at that time. In Brenda’s class with focus group D, the group’s model of resampling with replacement was contrasted to models of resampling without replacement. Discussion of focus group D’s models led all groups in the class to pursue resampling with replacement and Brenda did not think that the follow-up activity (discussed next) was needed [Instructor Interview, Day 6].

**Issues with resampling without replacement.** Although no groups in one of Allen’s classes initially constructed models of resampling with replacement, he discussed some possible issues with sampling without replacement. Allen discussed how the proportion of peanuts in the bag of nuts changed after each nut was drawn, but did not discuss the idea of resampling with
replacement to remedy this issue. Another issue Allen discussed with the class was what outcomes were possible when collecting a sample of mixed nuts from the population versus the possible outcomes when resampling without replacement. Cara realized that while it was possible to get any proportion of peanuts in a sample of mixed nuts, it was not the case using her approach of resampling. Her group collected resamples of 10 nuts without replacement.

  Allen: If I pluck 10, is it possible that I get all 10 peanuts?
  Cara: No, because there’s only 7.
  Allen: So I’ll never get 10 peanuts. I’ll never get all peanuts. Is it possible, in the world, if I have this massive barrel at [local grocery store], is it possible that they could give a 25-nut sample that had all peanuts?
  Class: Yeah.
  Allen: But we’ll never see that in the way we simulated it. [Class discussion, Mixed Nuts, Day 6]

This issue with resampling with replacement led into the follow-up class discussion to construct a way around this issue.

  *Resampling with a spinner vs. without replacement.* No groups in Allen’s classes initially constructed models of resampling with replacement. The non-focus group that used the spinner did construct a model that was structurally similar to resampling with replacement and would produce equivalent results. Allen discussed with his class how resampling with the spinner contrasted with sampling with replacement and how the bag of sticks could be used to produce equivalent results to the spinner. Allen and George (focus group B) discussed that when using the spinner, there is a 28% of choosing a peanut, just as drawing one nut from the bag of nuts. Once a nut is drawn, the percentage of peanuts in the bag then changes. George asserted that in order
Allen: Who can say what their approach was like in terms of what you guys did? Can you explain what a spin represents? Can you represent it in terms of what you guys did? You guys pulled out sticks.

George: Oh I know. They took the original percentage of the bag, which is 28%, and they just spin it as many trials as they want to get what they got.

Allen: So if they do a spin, what’s that like for you guys with your sticks?

George: Picking 10 out?

Allen: They do one spin. What’s it like for you guys with your sticks?

George: It’s like one trial.

Allen: Pulling one more stick.

George: Pulling one stick, right.

Allen: It’s kind of like pulling one stick. But then, doing what with that stick? ‘Cause let’s say I pull out a stick that’s a nut that’s not a peanut. What’s the proportion left in peanuts in the bag? Has it changed at all?

Class: Yes.

George: So you have to put it back?

Allen: If I wanted to be using that same 28% right, because you had said that the 28 came from the 7 out of 25? [Class discussion, Mixed Nuts, Day 6]

This discussion brought out the key idea of resampling with replacement in order to preserve the proportion of peanuts in the original sample. The class’ discussion then segued into the Trident Gum class discussion in order for the students to pursue these ideas of replacement and preserving the proportion in the original sample.
Resampling with replacement vs. without replacement. Like the previously discussed class where a group constructed the spinner, Brenda’s class with focus group D discussed the idea of preserving the proportion of peanuts when choosing each nut. In addition to preserving the proportion, Susan from focus group D discussed an idea related to the relationship between resampling with replacement and sampling from a population.

Brenda: How did you settle on that and how instead of doing what these other groups have done and grab 10. Why did you do it this way?

Susan: My thought was that if we pulled out the nut, we can’t repick it out of the bin but from a bigger perspective, you can’t really determine what’s it’s going to be like if you don’t put them back and have that same amount of like, I don’t know…

Brenda: So you’re talking about the reason why you put it back in.

Susan: Yeah. ‘Cause you wouldn’t have that big scale and have as much variety. [Class discussion, Mixed Nuts, Day 6]

Susan asserted that when taking a sample of nuts from the bin (or population), you couldn’t choose the same nut twice, but when resampling, that nut is put back to keep the percentage of peanuts the same, and represent another nut.

Follow up activity: Trident gum. A whole class discussion was planned to help elicit resampling with replacement and develop groups’ models of resampling towards the method of bootstrapping. The discussion involved an activity that was similar in structure to the Mixed Nuts MEA. When planning the second model development sequence, I anticipated that groups may struggle to construct ideas of bootstrapping and need additional experiences to elicit the concept. The activity was planned as a class discussion, rather than a group activity, in order to
save classroom time that would be used for the remaining MXA and MAA. Three classes, excluding Brenda’s class with focus group D, discussed the follow-up activity. In Brenda’s class with focus group D the Mixed Nuts MEA elicited ideas of resampling with replacement with two groups. After that class discussed those groups’ models of resampling with replacement, Brenda decided that the follow-up activity was not needed for that class.

Trident gum had an advertising campaign that claimed that four out of five dentists recommend Trident gum to their patients who chew gum. The question posed to the class was: If the class asked five dentists if they recommend Trident gum, what is the likely range of the five that would recommend chewing Trident gum to patients who chew gum? The discussion was approached in a similar manner in the three classes that pursued it, so I will only report the discussion and the development of ideas of resampling in Allen’s class with focus group B. Allen began by asking the class if they could simulate a group of five dentists using five note cards. The class decided that four of the cards could represent the four dentists who recommend the gum and the remaining card for the fifth dentist who does not. The class then discussed how to resample from the five cards.

Allen: How could I possibly use something like notecards to simulate 5 dentists and simulate the likely range of results?

Wilma: Write yes on four and no on one?

Allen: Could I write R for recommend?

Class: Yeah.

Allen: So I’ve got R, and there’s one dentist who thinks it should be, I don’t know, Bazooka, or something. So I shuffle and I shuffle and I say go ahead and get your four, or how many should he draw?
Wendy: Five?

Wilma: One?

Allen: If you take all five, you’re going to get the four, right? You’re going to get the four and the one.

Wendy: Three? [Class discussion, Mixed Nuts, Day 6]

As a class, three samples of three cards were collected without replacement from the five cards. This led to a similar discussion to the mixed nuts MEA about how it was not possible to collect all possible types of samples. Since there was only one card that represented a dentist that did not recommend Trident gum, none of the collected samples would have more than one non-recommending dentist.

Allen: Is it ever possible that we’re only going to get one out of three [dentists who approve]? Think what I’ve got here.

Wendy: It’s possible.

George: No it’s not, you only have one.

Allen: Can you get zero out of three? If you’re pulling three, can you get zero dentists [who approve]?

Class: No.

Allen: So your only options are you’re going to get the three or you’re going to get the two [that approve]. [Class discussion, Trident Gum, Day 6]

Just as in the mixed nuts MEA, when resampling without replacement, it is not always possible to collect all outcomes. In this Trident gum simulation, it may be more noticeable that this is an issue, since when collecting a resample of three cards there are only four outcomes (0, 1, 2, and 3 dentists recommending the gum). With the model of resampling constructed by the
class, half of the outcomes are not possible to be collected, while it is reasonable to assume that it is possible that there are three dentists where two or three do not recommend chewing Trident gum. The smaller sample of five dentists compared to 25 nuts may have contributed to highlighting the need for sampling with replacement.

This lead to discussion of a new model of resampling that addressed these issues by choosing five cards with replacement, to simulate a group of five dentists’ approval of Trident gum. Allen asked the class if they had any other ideas to simulate data.

Wesley: Draw one and put it back 5 times?

Allen: Okay, so you said draw one and put it back.

Allen: Does it matter if I put this back in or not?

Class: Yes.

Allen. Because that will skew it. [Class discussion, Mixed Nuts, Day 6]

Two students then chose two resamples of five dentists with replacement from the five notecards. Each sample had four out of five dentists recommending Trident. Allen then opened a TinkerPlots environment in the front of the class (Figure 44), which contained a sampler window with five balls, four marked “Y” for yes they recommend and one marked “N” for not recommended. Similarly to each previous TinkerPlots environment, there was also a results window to collect a list of resamples and a dotplot window.

Allen: Let’s say that I happen to have four out of five dentists. Four say yes and one says no. How many am I going to draw?

Wesley: Five.

Allen: Let’s get 5. [Allen uses TinkerPlots to collect one resample of five dentists]. Did you see what it did? It pulled out five cards.
Wesley: Four out of five!

Allen: But that’s only one dot. Do you think that’s the answer then?

George: No, repeat 100 times.

Allen: Okay, come up and adjust the thing [TinkerPlots to collect 100 samples]. Okay, are you happy with 100 samples?

Wesley: No, more.

Allen: What would you say based on these 100 simulations?

Wesley: Not enough data. [900 more samples were collected, Figure 44]

Allen: Are you starting to get a good sense of what the likely range of answers would be here?

Figure 44. TinkerPlots environment for Trident Gum projected in the front of the classroom collecting 1000 samples of five dentists. [Class discussion, Mixed Nuts, Day 6]

The class found that after collecting 1000 samples there was still no group of five dentists where none approved of chewing Trident gum, so they choose to collect 10,000 more samples for a total of 12,000. Allen then summarized what was done to simulate data and referenced it back to
the Mixed Nuts MEA.

Allen: We only had those four Rs [dentists who recommend Trident gum] out of five to draw from. How many did we draw? We drew five to try and simulate what could possibly come out. Sometimes it could come out that we drew that five straight times, that was pretty unlikely, right? Most of the time, it was a 3, 4, or 5. Do you have an idea now how we can go back to our little sample, the little 100 calorie pack [of mixed nuts]? How many did we draw with the cards?

Wendy: One.

Allen: Just one and we put it back in, shuffled it around again and pulled another. So who’s got a thought now on how we can approach this going back to our little 100-calorie pack [of mixed nuts]?

Wesley: So you take one nut, count that, and put it back in. Then take out another one.

Allen: How many of these should I draw to get a sense of what the bag might look like?

Wendy: 25.

Allen: 25. I’m going to draw essentially out the whole bag, but I’m not going to draw them all out at one. This is going to be a bit of a lengthy process. We’re not going to get a lot of these samples, but maybe you guys can do a couple and you guys do a couple and we can put them together and see where we’re at. [Class discussion, Mixed Nuts, Day 6]

Each group then generated two bootstrap samples of 25 mixed nuts using the marked craft sticks (seven peanuts, 18 other nuts), for a class total of eight resamples. The data was displayed on the white board in the front of the room (Figure 45). From this data the class determined that the likely range for the percentage of peanuts was 28% to 36% (7-9 peanuts out of 25 nuts).
Generating bootstrap samples with hands-on manipulatives can be a time demanding process. The next MXA in the second model development sequence introduced a TinkerPlots environment that performed bootstrapping nearly instantaneously.

![Aggregated class dotplot for bootstrap samples of mixed nuts collected with hands-on manipulatives.](image)

*Figure 45. Aggregated class dotplot for bootstrap samples of mixed nuts collected with hands-on manipulatives. [Class discussion, Mixed Nuts, Day 6]*

**Exploring models of resampling with technology.** A key aspect of using TinkerPlots is the ease with which bootstrap resamples can be taken and represented in a dotplot. In the previous Mixed Nuts MEA, groups that constructed models of resampling with replacement did not draw conclusions regarding the percentage of peanuts in the brand of mixed nuts. Focus group D collected only one resample in this way due to how much time it took to collect the sample. A non-focus group pursued methods of resampling without replacement rather than with replacement because they thought the process was inefficient because it took too long to perform.

In the next MXA, students used TinkerPlots to simulate data with bootstrapping and collected many samples nearly instantaneously. The goals of the MXA were to explore the model of bootstrapping to determine likely outcomes from a dotplot representation of the simulated data and to compare and reason with the structure of the distributions of simulated data that were generated by different original samples.

The MXA’s TinkerPlots environment consisted of a sampler containing 25 balls with seven marked as “P” for peanut and 18 as “O” for other types of nuts. The sampler selected one
ball at a time, with replacement, and chose a total of 25 balls for each bootstrap resample. Just as in the previous *TinkerPlots* environments, there were also results and dotplot windows to display the data. The MXA first guided students to simulate data for the sample of mixed nuts used in the MEA (seven peanuts out of 25 nuts) and predict the likely range for the percentage of peanuts in this brand of mixed nuts. Each student was then given a second sample of 25 nuts, containing one to 11 peanuts, and asked to use this new sample to simulate data and predict the likely range for the percentage of peanuts in this brand of mixed nuts. Students were instructed how to adjust the *TinkerPlots* sampler to coincide with the composition of their new samples of mixed nuts. Later in the activity, students were told that the manufacturers of the brand of peanuts had stated that the percentage of peanuts varies slightly for each batch of mixed nuts that is produced, but the current batch contains approximately 21% peanuts. Students were asked if the value of 21% peanuts was in their predicted ranges of values, and what did that tell them? Students used their predicted ranges to comment on the methods used to construct the predicted ranges, representativeness of their samples, use of their samples as evidence against the brand’s claim of 21%, and distribution of samples in the population of mixed nuts.

*Methods used to construct the predicted ranges.* Some students asserted that if 21% was or was not in their predicted ranges, it supported the accuracy or inaccuracy of the prediction. One non-focus group student found that 21% was in both of his predicted ranges and asserted that: “Yes, it tells me that our method was accurate” [Written classwork, Mixed Nuts with *TinkerPlots*, Day 7]. This would seem to imply that an accurate method would always produce a range of percentages that contains the actual percentage of peanuts in the brand of mixed nuts. This discounts the chances of collecting a sample of mixed nuts that is not representative of the population. In this case, an accurate procedure could produce a range of values that does not
capture the actual percentage of peanuts.

Representativeness of the sample. Another non-focus group student discussed ideas related to some predicted ranges being more accurate than others. She stated that:

21% was only in one of my likely percentages for peanuts. This would tell me that my second simulation w/ 5 peanuts [in the sample of 25 nuts] and a range of 3-6 [peanuts per 25 nuts] would be more accurate. [Written classwork, Mixed Nuts with TinkerPlots, Day 7]

More accurate in this case does not seem to apply to the procedure to construct the range, but the sample. A sample that is a more accurate representation of the population would construct a predicted range that is more likely to capture the actual percentage of peanuts in the brand of mixed nuts. This MXA was the first activity in the instructional unit where individual students were using different samples or populations to simulate data. In the Ophelia MEA and MXA, all simulations were based on her 50-50 chance of guessing the winner of the basketball games. In the Nike sneaker MXA and MAAs, each group used a bin of beads with an identical composition of beads. For the first mixed nuts MEA, all groups received a sample of 25 nuts with seven peanuts. In Mixed Nuts MXA, each individual received a sample of 25 mixed nuts with one to 11 peanuts, which led Randy (from focus group D) to emphasize, “the number of peanuts in each sample has an impact on determining the likely range” [Written classwork, Mixed Nuts with TinkerPlots, Day 7].

When learning that the actual percentage of peanuts in the brand was 21%, some students discussed the representativeness of their sample and how it related to other samples. Megan (focus group C) stated that 21% “was not in my predicted range, but it was close. So it tells me … that my one sample does not fully describe or is even accurate to the actual percentage”
By using the words “fully describe” and “accurate,” I think that Megan is asserting that her sample is not representative of the population since she used the sample to construct a likely range that did not contain the actual percentage of peanuts. George (focus group B) also discussed ideas of a representative sample and suggested that since his sample was higher than the actual 21%, some samples may just have a higher percentage of peanuts. George stated that 21% “seems lower than what my bag was showing that mixed nuts bags, some may contain more peanuts than others” [Written classwork, Mixed Nuts with TinkerPlots, Day 7]. A non-focus group student also discussed the distribution of samples and stated, “my two ranges were 30-50% (7 peanut trial) and 32-44% (9 peanut trial). This shows that there were more peanuts in my trials than expected” [Written classwork, Mixed Nuts with TinkerPlots, Day 7]. She noted that she expects that if the actual percentage of peanuts is 21%, that’s what she would expect to collect in a sample.

**Predicted ranges as evidence.** Students differed in how they used their data to draw claims. One non-focus group student constructed a likely range that did not include 21% and concluded that the company’s claim of 21% did not agree with his data. When asked if 21% was in the student’s likely range, he stated “No, my likely range was 4-16% so the company’s approximate guess of 21% seemed high compared to my samples. This tells me that somebody’s wrong” [Written classwork, Mixed Nuts with TinkerPlots, Day 7]. He was unsure if his data showed that the brand did not have 21% peanuts in its mixed nuts or if his prediction was incorrect. Another non-focus group student took this reasoning a step further and claimed that the brand’s claim of 21% peanuts must be incorrect. The student stated that: “21% was slightly below my initial 24% prediction, and this tells me that 21% may not be a solid result, and that it can fluctuate with each batch” [Written classwork, Mixed Nuts with TinkerPlots, Day 7].
initial prediction, I interpreted that the students meant the lower endpoint of his likely range for the percentage of peanuts in the mixed nuts. The question in the worksheet stated that the percentage of peanuts was approximately 21% but varied between batches. This student asserted that since his sample produced a likely range above 21%, the sample may have come from a batch of mixed nuts with more than 21% peanuts.

Distribution of the population. Lena (focus group C) used her predicted ranges to determine what she thought would be a representative sample of mixed nuts. The likely ranges for the percentage of peanuts that she predicted were above and then below the actual percentage of 21%. This led her to determine that a representative sample would contain a number of peanuts greater than her sample that produced a range below 21% and less than her sample that produced a range above 21%. Lena stated:

21% peanuts was not in either of the predicted ranges for percentage of peanuts. For 7/25 peanuts the range was 25% - 32% and for 3/25 peanuts the range was 12% - 16%. This data tells me that in order to get 21% peanuts the number of peanuts has to be higher than 3 and less than 7 in a sample of 25 nuts. [Written classwork, Mixed Nuts with TinkerPlots, Day 7]

By claiming that samples of 25 nuts containing more than 3 and less than 7 peanuts are representative samples, Lena was describing the likely range of peanuts in the sampling distribution constructed when repeatedly sampling from the population of mixed nuts.

Comparing resampling distributions. The final activity of the second model development sequence was an MAA that was similar to the final MAA of the first model development sequence and asked groups of students to compare two distributions of data. In the final MAA of the first model development sequence, two different characteristics (Nike and
Adidas) were compared in the same population. In this MAA, the same characteristic (peanuts) is compared in two different populations. The goal of the MAA was to apply the model of bootstrapping to compare two populations and use simulated data to make claims about the populations from which the data was drawn. Each group was given two samples of mixed nuts, with the two samples coming from different brands of mixed nuts. The groups were told that the manager of the local grocery store planned to order a large shipment of mixed nuts and wanted to determine from the two samples which brand was likely to have the smaller percentage of peanuts. Each group was given a different pair of samples. Each pair of samples contained two samples of 25 nuts with the differences in peanuts between the samples varying from two to eight peanuts. I chose these differences so that I could determine what differences groups viewed as evidence to select one brand of mixed nuts over the other, and also so that groups could observe the varying differences in the likely ranges constructed during class discussion of the MAA. Groups had access to the TinkerPlots environment from the Mixed Nuts MXA and were shown in the previous MXA how to change the composition of the nuts in the TinkerPlots sampler window to align with their samples.

Groups applied their models of resampling and inference to respond to the manager in three ways. For the first response (n=5, including focus group D), the likely ranges for the percentage of peanuts in each brand did not overlap, leading the groups to conclude that the store manager should purchase the brand which produced the lower range of values. In the second response (n=7, including focus group C), the likely ranges for the percentage of peanuts in each brand did overlap, but the groups still concluded that the store manager should purchase the brand which produced the lower range of values. For the final response (n=4, including focus groups A and B), the likely ranges for the percentage of peanuts in each brand also overlapped,
but the groups noted some uncertainty about which brand of mixed nuts contained fewer peanuts.

_No overlap._ Groups constructing predicted ranges that did not overlap had pairs of samples with the following proportions of peanuts to mixed nuts: (5/25, 1/25), (14/25, 6/25), (11/25, 5/25), and (12/25, 5/25). Focus group D used each of their samples (14/25, 6/25) to simulate 100 bootstrap resamples in _TinkerPlots_ and determined that the likely range of peanuts in each brand of mixed nuts were 48-66% and 16-28%, respectively. Focus group D determined that:

After looking at our two random samples of choice A and B the better choice would be brand B. After conducting 100 trials based off of Brand A and Brand B the likely range of peanuts in brand A is 48% to 64% and in brand B the likely range is 16% to 28% making brand B the best choice because it has less peanuts. [Written classwork, Comparing Mixed Nuts, Day 8]

They asserted that because their likely ranges for the percentage of peanuts did not overlap, that implies that brand B “has” fewer peanuts.

The non-focus groups giving this form of response varied between using language of certainty and uncertainty when making their claims. A non-focus group with the samples 5/25 and 1/25 peanuts to mixed nuts constructed likely ranges for the number of peanuts in samples of 25 and stated, “Our range from the data for B is 0-2 for A it is 3-6. Clearly B has less peanuts because A will never be lower than 3” [Written classwork, Comparing Mixed Nuts, Day 8]. By saying that A will never go below 3, they asserted that there is no possibility that brand A has fewer than three peanuts per 25 nuts, rather than stating that it is unlikely. Another non-focus group combined language of certainty and uncertainty to assert: “We decided to go with brand B because overall the average was lower and it didn’t have an overlap. It would have less peanuts
in it and most of the time it would be in that range” [Written classwork, Comparing Mixed Nuts, Day 8]. They spoke of uncertainty by claiming most of the time it would be in that range. I interpreted that this was referring to the percentage of peanuts in the brand of mixed nuts. This would suggest that while it is likely that brand B has fewer peanuts than brand A, it is possible but unlikely, that brand A has fewer peanuts.

*Overlap and chose lower interval.* Groups (n=7, including focus group C) giving this response constructed likely ranges for the percentages of peanuts in each brand of mixed nuts that overlapped. The groups recommended that the manager of the grocery store purchase the brand with the lower range of peanuts. Focus group C was given samples that contained 10/25 and 6/25 peanuts to total nuts. The group used *TinkerPlots* to simulate 50 resamples for each brand of nuts. The *TinkerPlots* environment only displayed one dotplot at a time, so Lena (focus group C) copied the dotplots onto the activity worksheet to view them both at the same time (Figure 46).
Figure 46. Lena’s (focus group C) dotplots of simulated data for the number of peanuts in bootstrap samples of 25 mixed nuts for each brand of mixed nuts. [Written classwork, Comparing Brands of Mixed Nuts, Day 8]

From the dotplots, focus group C determined that the likely range for the percentage of peanuts in brand A was 32%–44% and for brand B 20%–32%. The group concluded, “she [the store manager] should get B because it is a smaller range of 20%–32% of peanuts compared to 32%–44% of peanuts” [Written classwork, Comparing Brands of Mixed Nuts, Day 8]. The group did not address the overlap in their likely ranges for the percentage of peanuts and only focused on the range with the lower percentages on peanuts.

Nate (focus group C) created a visual representation of the simulated data that was different than the dotplots (Figure 46) created by Lena (also focus group C) Nate copied both dotplots shown in TinkerPlots onto his worksheet, but combined the data onto one set of axes. Each data point simulated from brand A’s sample was plotted as an A and each data point from brand B was plotted as a B (Figure 47).
Figure 47. Nate’s (focus group C) combined dotplot for simulated data for the number of peanuts in bootstrap samples of 25 mixed nuts for each brand of mixed nuts. [Written classwork, Comparing Brands of Mixed Nuts, Day 8]

The dotplot (Figure 47) shows evidence that the resamples with more peanuts tended to come from the sample from brand A and the resamples with fewer peanuts tended to come from the sample from brand B.

Three groups giving this response constructed likely ranges for the percentage of peanuts in brands A and B that overlapped only at the endpoints. The remaining three of the seven groups, constructed likely ranges with more than only the endpoints overlapping. One non-focus started with samples having proportions of 8/25 and 5/25 peanuts to total nuts from brands A and B, respectively. The group used TinkerPlots to generate 1000 resamples for each brand of nuts and from the dotplots concluded that the likely range of peanuts in brand A was 24%-44% and for brand B was 12-28%. Just as with focus group C, this non-focus group did not discuss the overlapping values in their conclusion, and discussed the lower number of peanuts in their conclusion. The group stated that:
I would choose bag B with 5 peanuts. The range is 12%-28%, which is pretty low. They have a likely chance of choosing a lower number of peanuts because of the range, also the numbers on the dotplot on the lower end have a more likely chance of being chosen.

[Written classwork, Comparing Brands of Mixed Nuts, Day 8]

This non-focus group discussed ideas of uncertainty in their response. Their likely range of values for brand B contained values lower than the range for brand A, which made the group assert that it is more likely that brand B would contain the lower percentages of peanuts than brand A. The lower percentages were in the likely ranges for brand B, while below the likely range for A. Just as all groups giving this form of response, there was no mention of the possibility that since the brands had overlapping likely ranges for the percentage of peanuts, it would not be unlikely for the sample with the higher proportion of peanuts to come from the brand that actually has the lower percentage of peanuts.

**Overlap and uncertainty over choice.** In the final form of response, four groups (including focus groups A and B) constructed likely ranges for the percentages of peanuts in each brand of mixed nuts that overlapped. Each group discussed in their conclusion an uncertainty over which brand the store manager should choose because of this overlap. Focus group A began with samples of brand A and B that contained proportions of 6/25 and 2/25 peanuts to total nuts, respectively. The group used *TinkerPlots* to simulate 1000 bootstrap resamples for each brand of mixed nuts. From the dotplots, the group concluded that the likely range for the percentage of peanuts in brand A was 16% to 32% and 0% to 16% for brand B. The group determined:

Because 16% is found in both samples’ likely ranges, it is hard to give you an accurate answer to your question of which sample has a smaller percentage of peanuts. You could have a likely result of 4/25, 16% peanuts from each sample. It is really luck of the draw.
However, we would like to be as helpful as possible, therefore, we would advise you to choose sample B because overall, there is a smaller percentage of peanuts that we observed. [Written Classwork, Comparing Brands of Mixed Nuts, Day 8]

The group referred to the sample many times in their conclusion. I interpreted this to mean the population from which the sample was collected. The MAA asked which brand of mixed nuts had the lower percentage of peanuts. This refers to the population of mixed nuts, not the sample as written in focus group A’s conclusion. The group also stated, “you could get a likely result of 4/25, 16% peanuts from each sample.” I interpreted that they are referring to the likely proportion in the population of each brand, which they predicted based on their samples.

Focus group A asserted that because their likely ranges for the percentage of peanuts for brand A and B overlapped, they were not able to give an accurate response regarding which brand has less peanuts. By accurate, I interpreted that the group meant that since there is an overlap, it would not be unlikely for each brand of peanuts to contain 16% peanuts. Because this would be a likely result, the group could not choose accurately between the two brands. After prefacing that their recommendation may not be accurate, focus group A did chose to recommend brand B since most of the range of likely percentages of peanuts was below the range for brand A. This reasoning seems to be similar to responses from the previous category of responses where groups chose the lower range of values without regard to the overlap.

A non-focus group came to a similar conclusion as focus group A. The group constructed likely ranges for the percentages of peanuts that overlapped at the endpoints. The group asserted that the evidence was not convincing to choose one brand over the other, but they also considered the brand producing the lower interval as the better choice since it was more likely to have a lower percentage of peanuts, given their samples. This non-focus group concluded that:
Our data overlaps a bit … since sample B ends with 32% and sample A begins with 32% they overlap. Sample B will most likely be the best choice due to the lower percentage, but the evidence is not convincing. [Written classwork, Comparing Brands of Mixed Nuts, Day 8]

Focus group A gave a recommendation that was prefaced as being inaccurate. This non-focus group’s response differed in that they cited their evidence as unconvincing, rather than inaccurate, and chose the brand that they thought was more likely to have the lower percentage of peanuts.

Focus group B drew the conclusion that an overlap in likely ranges made them unable to accurately recommend one brand over the other, but unlike focus group A, did not choose brand A or B to recommend to the store manager. Focus group B started with samples for brands A and B containing 7/25 and 9/25 proportions of peanuts to mixed nuts in their samples, respectively. The group used TinkerPlots to simulate 1000 resamples for each brand of mixed nuts and determined that the likely ranges for the percentage of peanuts in brand A was 10%-36%, and for brand B was 28%-40%. This overlap of eight percent was more of an overlap than for the ranges constructed by focus group A where only the endpoints of the intervals overlapped. Focus group B concluded that: “due to the overlap of the ranges it is unclear to say which brand really has the smaller percentage of peanuts” [Written classwork, Comparing Brands of Mixed Nuts, Day 8].

The remaining non-focus group in this category of response was similar to focus group B’s response, and asserted that they were unable to recommend one brand over the other. This non-focus group began with very similar samples from brands A and B, 9/25 and 8/25 proportions of peanuts to nuts, respectively. The group constructed intervals that overlapped for the majority of the intervals (20%-52% and 16%-44%) and determined that since the samples
were so close together, they could not choose one brand over the other.

**Summary of Model Development**

Throughout this chapter, I have reported the models of sampling and inference that were elicited and developed by groups of students while participating in the instructional unit. The construction of the instructional unit was driven by investigating three types of students’ reasoning:

1) Students’ reasoning that developed as they moved from repeated sampling methods to bootstrapping methods

2) Students’ reasoning that developed about the method of bootstrapping in order to make informal inferences about a population of data.

3) Students’ reasoning that was revealed and supported by moving from using hands-on manipulatives to computer simulations during repeated sampling and bootstrapping activities.

For the remainder of this chapter, I summarize the results of students’ model development over the course of participating in the instructional unit as it applies to these three types of students’ reasoning.

**Moving from repeated sampling to bootstrapping.** In the first model development sequence, models of sampling and inference were elicited in which the groups used both hands-on manipulatives and *TinkerPlots* to simulate data. Groups of students organized this data in tables and dotplots, and developed models to examine these representations of the data to make inferential claims regarding how likely it was for various outcomes to occur.

From these representations, groups of students determined that the outcomes that occurred most often in their simulated data were likely to occur most often in the population.
Students found that one key to this correspondence was the collection of many samples (or trials or resamples). When simulating a small number of samples, the empirical sampling distribution may not be symmetric, with the outcomes occurring most often not in the center of the distribution. When simulating many samples, the distribution approached a bell-shaped curve.

Focus group C determined in the first model development sequence that the sampling distribution was bell curve shaped. This group used this aspect of the distribution to anticipate the center of the sampling distribution without collecting many samples. TinkerPlots allowed the students to collect as many samples as they liked, but focus group C persisted throughout both model development sequences to collect fewer samples (approximately 50 to 150) than many groups who tended to collect more than 1000 samples. Focus group C anticipated the shape and center of the distribution from these fewer samples and determined that it was not necessary to collect more samples in order to make their predictions for the likely range of outcomes.

Some students also found that the collection of many samples increased the consistency between group members’ individually constructed empirical sampling distributions. When each student collected small numbers of samples, individual group members’ distributions looked different from one another’s, but as more samples were collected, the distributions began to look more similar. This led students to determine that collecting more samples would lead to a consistency between the inferential claims made with different sets of simulated data. There was not a consensus from groups as to what number of samples constituted many or enough samples. Some students, including those in focus groups A and B, found that as the number of samples increased beyond a certain number (approximately 1500), their empirical sampling distributions’ shapes barely changed. Other students, including those in focus groups C and D, felt that after collecting approximately 100 samples, they could feel confident with their inferential claims
Groups of students applied their ideas of collecting many samples in the second MEA that used hands-on manipulatives to elicit bootstrapping methods. For some groups of students, including focus group D, resampling with replacement was elicited by the MEA, but students viewed the process as inefficient. They instead chose methods that resampled without replacement that would more quickly collect enough samples for the group to make inferences about the population. This choice allowed the groups to construct many samples in a reasonable amount of time. In the first model development sequence, 20 samples could fairly quickly be generated with hands-on manipulatives by flipping eight coins 20 times or using the spatula to scoop 20 samples of 20 beads.

**Bootstrapping methods.** A key aspect of the bootstrapping method is the assumption that the sample is representative of the population. When considering the sampling distribution for samples collected from the population, it is possible although unlikely to collect a sample from the extreme ends of this sampling distribution. It is more likely to collect a sample from the central region of the sampling distribution, which would be representative of the population. The second key aspect of bootstrapping is a process of resampling that preserves the composition of the representative sample while drawing each element of the bootstrap sample. This is accomplished through resampling from the sample with replacement. During the second model development sequence groups of students investigated these two key aspects of bootstrapping. Models of resampling and inference were elicited with these two aspects in mind.

*Representativeness of the sample.** In the Mixed nuts MEA, groups of students had only one sample to draw inferential claims about the population. Students from focus group D asserted that while this sample may not have a composition of peanuts that is an accurate
representation of the population, it is the only sample that they had to work with. There is the possibility that the one sample was very different from other samples taken from the same population, but it is more likely to be similar to other samples. In the Mixed Nuts MXA, groups of students compared their likely ranges of values to a parameter in the population. Some students determined that if a sample was an accurate representation of the population, then they would construct a predicted range of values that is more likely to capture the actual parameter in the population. Students from focus group C determined that because their likely range of values did not capture the actual parameter in the population, their original sample must not be representative of the population.

*Resampling with replacement.* The Mixed Nuts MEA elicited some groups of students, including focus group D, to construct models of resampling with replacement (or equivalent) from the initial sample of mixed nuts. Two non-focus groups that initially constructed a model of resampling with replacement decided to pursue different methods, in large part due to the time demanding nature of using hands-on manipulatives to resample with replacement. Focus group D pursued this method, but ran out of class time and only had time to collect one resample. The Mixed Nuts MEA elicited a fourth non-focus group of students to construct a model of resampling that would produce equivalent results to resampling process with replacement. This group built a spinner out of a pencil and hairpin with areas equivalent to the proportion of peanuts and other nuts in the sample. By creating the spinner, this group constructed a resampling process that upheld the composition of the sample when selecting each element of the resample.

No groups initially constructed the method of bootstrapping, since the sizes of the resamples taken within each group were not equal in size to the original sample. One reason for
collecting smaller sized resamples may have been the time demanding nature of the process. Another possible reason is that a similar method was used in Nike Market Share MAA to collect samples of 20 sneaker purchases from a population of 4000. Students collected samples much smaller in size compared to the population in order to make claims about the population. Students may have drawn a corollary between these two activities and thought that it was a reasonable process to use to make claims about the percentage of peanuts in the population of the mixed nuts.

**Moving from hands-on manipulatives to technology.** The use of TinkerPlots allowed students to see the structure of empirical sampling distributions that was best observed after collecting large numbers of samples. In the activities that used hands-on manipulatives, groups constructed dotplots containing no more than 20 samples. Aggregating the data from all groups in a class led to students observing the structure more clearly, but not as clearly as when thousands of samples were quickly collected in TinkerPlots. By collecting large numbers of samples, students determined that (in most cases) the empirical sampling distribution was bell curved, pyramid, or mountain shaped. In the last two activities of the instructional unit, the symmetry did not exist for students who had samples with proportions of peanuts close to zero. The dynamic nature of TinkerPlots’ dotplots allowed students to observe how the shape of the dotplots changed as more samples were collected. When 10 samples were collected, students found that the shape of distribution was hard to decipher. After 50 samples, the distribution began to have an approximate mountain shape. With 1000 samples, the distribution looked like a bell-shaped curve and the collection of more samples had little influence on this shape. Viewing this process of the empirical sampling distribution grow as more samples were collected allowed students to anticipate the shape of the distribution without actually collecting more samples.
There was not a consensus among students regarding how many samples should be collected to make inferential claims. Those students collecting 1000 or more samples reasoned that they needed this many samples to construct an empirical sampling distribution that was shaped like a bell curve. Students who thought that fewer samples, such as 100, were required to make claims may have thought that they could view the dotplot and anticipate what the plot would look like if many more samples were collected. Once the center of the distribution was located, students then determined a likely range of outcomes from their anticipated sampling distribution. This anticipation of the distribution’s shape could be very useful reasoning for students’ when simulating data with time-consuming processes such as using hands-on manipulatives.
Chapter 5 – Discussion and Conclusions

The findings of this study contribute to the field of statistics education by examining students’ reasoning as they constructed and developed bootstrapping methods and investigating the relationship between this reasoning and the drawing of informal inferences. Students participated in a four-week instructional unit that was designed as two model development sequences. The first model development sequence elicited models from groups of students which were used to repeatedly collect samples from a population in order to make inferential claims from empirical sampling distributions. In the second model development sequence, students had only one sample from the population, without the option of collecting additional samples. From this one sample, groups of students constructed and developed models of resampling in order to make inferential claims from empirical resampling distributions. The model eliciting activity in the second model development sequence elicited models from some groups of students which were used to resample with replacement, although those models fell short of the bootstrapping process by not collecting resamples that were equal in size to the original sample. Class discussion of a follow-up activity, similar in structure to the model eliciting activity, encouraged students to consider the value of drawing resamples of equal size to the sample, which led students to construct the method of bootstrapping.

With the trend in statistics education moving from a focus on theoretical distributions towards the simulation and analysis of distributions of data, this study has implications for the design of future introductory statistics curricula. In this study, groups of students engaged in a model eliciting activity for which there was not one clear method for determining a solution. Through this activity and a follow-up discussion, students constructed bootstrapping methods in order to determine a means to answer the questions posed in the activity. Some research has
examined curricula for introductory statistics courses that emphasize data simulation and the bootstrapping process (Garfield, delMas, & Zieffler, 2012; Pfannkuch, Forbes, Harraway, Budgett, & Wild, 2013), but the key difference with this research is the *elicitation* of bootstrapping methods by groups of students, rather than the instruction of students on how to use the method. Research has indicated that the elicitation and development of models has led to significant forms of learning (Lesh et al., 2000).

This study was guided by three research questions which investigated students’ reasoning of sampling and inference with repeated sampling and bootstrapping methods.

1) What student reasoning develops as they move from repeated sampling methods to bootstrapping methods?

2) How do students develop their reasoning about bootstrapping methods in order to make informal inferences about a population of data?

3) What student reasoning is revealed and supported by moving from using hands-on manipulatives to computer simulations during repeated sampling and bootstrapping activities?

In the sections that follow, I discuss my conclusions for each of the research questions and discuss my findings in terms of the research literature.

**Reasoning Developed Moving from Repeated Sampling to Bootstrapping Methods**

In the first model development sequence, groups of students constructed and developed their distributional and inferential reasoning. In the literature review, I synthesized research on distributional reasoning (Bakker & Gravemeijer, 2004; Friel, O’Connor, & Mamer, 2006; Reading & Reid; 2006) to include four hierarchical levels of distributional reasoning:
1) Reasoning with individual values in a data set.
2) Reasoning with a single aspect of a distribution.
3) Reasoning with multiple aspects, but not relating an understanding of the aspects.
4) Reasoning with the relationships between multiple aspects of a distribution.

Most groups, including all focus groups, developed distributional reasoning to reason with multiple aspects of one empirical sampling distributions (levels three and four), with focus on the center and spread of the distribution. During the first model development sequence, some groups (including focus groups B, C, and D) had a tendency to reason with only the center of a distribution. This is in line with research from Noll and Shaughnessy (2012) which discussed a persistence for student to reason with only the center of empirical sampling distributions. Throughout the model development sequences, these focus groups developed their models of sampling and inference to make inferential claims that took into account multiple aspects of the distribution. In line with Makar and Rubin’s (2009) central principles of informal inferential reasoning, the groups of students used probabilistic language, such as likely or unlikely, to reason with aspects of the data’s distributions in order to make generalizations beyond their collected data. The reasoning demonstrated by the groups drew parallels to the use of “modal clumps” (Konold et al., 2002), which were graphical interpretations of a distribution that focused on a range of values centered around the mean or median containing a high proportion of the data points. Most groups, including all focus groups, of students used this reasoning to construct likely ranges of outcomes with the center and spread of the distribution to make inferential claims. Groups of students determined that these likely ranges of values, or “modal clumps”, in their simulated data were likely to occur most often in the population.

This is in contrast to previous research (Saldanha & Thompson, 2002), which found that a
majority of the students did not focus on the distribution of sample statistics for inference and instead compared a single sample statistic with a population parameter. Saldanha and Thompson (2002) characterized their students who focused on multiple aspects of the distribution as having a multiplicative conception of the sample. This conception allowed students to view the sample as a “quasi-proportional, mini version” (p. 266) of the population, which can be used as to approximate the distribution of the population. Since the majority of participants in this study developed models to reason with multiple aspects of an empirical sampling distribution, I posit that they constructed this multiplicative conception of the sample, which was key to constructing and developing methods of bootstrapping.

**Reasoning Developed about Bootstrapping Methods**

The key difference between this study and previous research addressing data simulation and bootstrapping (Garfield, delMas, & Zieffler, 2012; Pfannkuch, Forbes, Harraway, Budgett, & Wild, 2013), is the *elicitation* of bootstrapping methods by groups of students rather than the instruction of students on how to use the method. Four categories of models for resampling and inference were constructed by the groups of students in the MEA of the second model development sequence (see Table 4).

1) Groups collected resamples from the sample, *without* replacement, with *consistent* sample sizes smaller than the original sample (including focus groups A, B, and C).

2) Groups collected resamples from the sample, *without* replacement, with *varying* sample sizes smaller than the original sample.

3) Groups collected resamples from the sample, *with* replacement, with *consistent* sample sizes smaller than the original sample (including focus group D).

4) Groups created a spinner with areas proportional to the composition of the sample
and collected resamples, with consistent sample sizes smaller than the original sample.

Groups constructing the first two models of resampling applied a partial multiplicative conception of the samples (Saldanha & Thompson, 2002) by treating the samples as representative of the population and resampling from the sample in the same way they had sampled from the population in the first model development sequence. What was missing from these models was the quasi-proportional nature of the relationship between the samples and population and how this relationship is affected by collecting samples without replacement. In the Mixed Nuts MEA, as each stick was drawn from the bag, without replacement, the proportions of peanuts changed, which meant the sample no longer maintained the representative relationship with the sample. Groups constructing the last two models of resampling took into account how the quasi-proportional relationship between the sample and population was affected by resampling and developed models of resampling which maintained this relationship.

Reasoning Moving from Hands-on Manipulatives to Computer Simulations

When working with hands-on manipulatives, students collected their data in tables. After all samples were gathered, many groups then constructed visual representations of data. A main difference between the visual representations created from the data simulated with hands-on manipulatives compared to the TinkerPlots activities was the static nature of these representations. When collecting samples in TinkerPlots, the representations were dynamic, with each sample added to the dotplot as it was collected. TinkerPlots allowed the students to observe the collection of samples in a visual, dynamic, and interactive manner with the students engaging in experimentation with data (Olive & Makar, 2010). I posit that the use of TinkerPlots to simulate the data was key to students’ reasoning with multiple aspects of one distribution. The
*TinkerPlots* environments allowed the students to investigate how distributions of data changed as more samples were collected. By observing the empirical sampling distribution grow in *TinkerPlots* as more samples were collected, students focused on observing the changing shape of the distribution as it grew to a bell curve. Research has described this focus on the distribution as a growing number of samples are collected as a global view of the randomness of sampling (Pratt & Ross, 2002).

**Implications for Future Research**

Three implications for future research arise from this study. They suggest further investigation of:

1. What environments would students design to simulate and represent data and quantities with the bootstrapping method?

2. How do students develop formal inferential reasoning after engaging in an instructional unit focused on data simulation and elicitation of bootstrapping methods?

3. What are teachers’ reasoning about data simulation and bootstrapping methods and the relationship between this reasoning and how these topics are taught in their classrooms?

**Student constructed sampling environments.** One future line of research could be to examine the representational spaces that learners create to illustrate the process of resampling, specifically bootstrapping. In each of the MXAs, I created the *TinkerPlots* environments to include my choice of sampler to simulate the data, a results window to display the data in a table, and a dotplot to visually represent the data. I made this choice over concerns for the learning curve for using new technology and the class time that it may have taken to instruct the students on how to create their own environments. In doing so, I made some choices for the students on the representations that they used in their models. While a focus of this study was elicitation and
development of students’ reasoning of sampling and inference when simulating data with technology, *TinkerPlots* was used to explore the structure of elicited models in MXAs rather eliciting students’ reasoning as in the MEAs. Because of this role of the software, I viewed it was a reasonable trade-off to construct the environment myself rather than investigate students reasoning as their models were elicited in *TinkerPlots*. If students had prior instruction with the *TinkerPlots* software before participating in the instructional unit, future research could investigate the manipulatives and representations within the *TinkerPlots* environment that student would use to construct and develop their models of sampling and inference.

**Development of formal inferential reasoning.** This study investigated how students develop reasoning of bootstrapping and informal inference. Informal inference has been viewed by researchers as a possible building block of formal inferential reasoning, which suggests that a future line of research is to investigate students’ development of formal inferential reasoning after participating in an instructional unit as implemented in this study. During the first model development sequence, students informally investigated situations where the formal method of hypothesis testing could have been applied, such as the Paul the Octopus Controversy homework assignment, and the first Nike sneakers MAA. The second model development sequence related to the construction of confidence intervals. Students used their one sample of mixed nuts to simulate the sampling distribution for the percentage of peanuts in samples of 25 mixed nuts. This distribution was then used to produce a range of likely values for the percentage of peanuts in the population of mixed nuts. Future research could investigate the connections that students see between the formal and informal methods conducted in this study and how the simulation of data with bootstrapping affects their future understandings of formal inference.
**Teachers’ reasoning.** The focus of this study was the development of students’ reasoning, although the instructors’ developing knowledge of data simulation and bootstrapping likely played a role in the students’ development. I designed the activities that were implemented in the classrooms, but it was the instructors who chose how to present the activities, choose the elements of simulation and bootstrapping to focus on during group and whole class discussions, and evaluated and gave students feedback on their written classwork. The nature of these interactions with the students would likely depend on the instructors’ own reasoning of simulation and bootstrapping. The Mathematical Education of Teachers II (Conference Board of the Mathematical Sciences, 2012) reported the need to implement statistics content-based professional development that focuses on the collection, analysis, and interpretation of data. Teachers prepared before the era of *The Common Core State Standards for Mathematics* (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010) may not have been exposed to data driven statistics and need opportunities to develop reasoning for these new methods.

**Limitations**

Two main limitations of the study involved the design and implementation of the instructional unit. First, everyone involved in the study, including myself, was new to teaching and learning through a modeling approach. Teaching and learning through a modeling approach takes time and experience to master. Second, rather than allow the students to freely use the *TinkerPlots* software to design models of their choosing, I designed the *TinkerPlots* environment. This may have limited the forms of models of sampling and inference that the students were able to construct.
**Initial experiences with modeling.** This study was the first experience in modeling for me, the instructors, and the participants. Lack of previous experiences with modeling affected the study at all levels, including my construction of the unit, the instructors’ implementation of the unit, and the students’ engagement in the unit.

*Researcher level.* I designed the MEA in the second model development sequence to elicit the method of bootstrapping. The activity elicited some models which groups used to resample with replacement from the sample of 25 nuts, but no groups collected resamples that were the same size as the original sample (until after Trident Gum class discussion). Collecting resamples of the same size is key in order to observe the variability in a statistic for samples like the original. While designing the MEA, I did not foresee issues with how the context of the activity encouraged students to make claims about the populations of mixed nuts, rather than the proportion of peanuts in future samples from the population of mixed nuts.

The context of the MEA used a bulk food bin of mixed nuts at the grocery store and asked students to “determine a likely range for the percentage of peanuts in the entire brand of mixed nuts” (Appendix C). The activity could be changed to work with small packages of mixed nuts, rather than bulk food bins, and ask the students to determine a likely range for the percentage of peanuts in each small package of mixed nuts for the brand. This change may refocus the need to examine samples of 25 nuts and elicit the method of bootstrapping.

*Instructor level.* Neither of the instructors had taught a model development sequence before, so it was not unexpected to have situations occasionally occur in the classroom that were not ideal for the development of students’ models. One instance occurred in the MEA of the first model development sequence. In this activity, groups constructed two models for how the coins were being used to simulate data. Some students viewed the sides of the coin as Ophelia’s correct
or incorrect prediction for the winners of a basketball game. Other groups chose the sides of the coin to represent whether the local university’s team won or lost the game. In each class where this divide occurred, the instructors asserted the correctness of models using the coins to represent correct versus incorrect predictions. Ideally, the students would have continued to pursue both models until the students saw the need to refine their model. This could also be considered a researcher-level issue since because I did not expect the students to represent the coins in this manner and did not discuss the possibility with the instructor prior to the MEA’s implementation.

Student level. The students were likely not accustomed to creating the type of final products that are meant to be created in MEAs and MAAs. The MEA in the first model development sequence instructed the groups to “write a letter to the [local newspaper] with the range of outcomes that you think are likely for Ophelia to correctly predict. Include a description of the methods that you used to come to this conclusion” (Appendix C). The letters tended to be directed to the instructors who would already have been familiar with how the groups had approached the activities, rather the newspaper staff which may not have had an understanding of data simulation. As the instructional unit progressed, groups did become more explicit in their letters and explanations for their recommendations and methodologies. Prior experiences in modeling may have helped the students to hit the ground running and begin the unit creating final products of this caliber. This again could be considered a researcher-level issue. I did not anticipate how explicit students should be when writing these letters. If I had, I may have been able to discuss with the instructors the types of letters that they should encourage their students to write.
Predesigned *TinkerPlots* environments. As I discussed in my implications for future research, I constructed the *TinkerPlots* environment for students rather than investing the class time that likely would have been needed for students to learn how to use the software to create their own sampling environments. In the first Ophelia MXA for instance, I made the choice to use a spinner with two equal regions to simulate her guesses for the winners of each basketball game, which was then spun eight times. *TinkerPlots* would have allowed the students to create different samplers to simulate this data, such as constructing eight spinners each to be spun once, mixers filled with balls as used in the later MXAs, or bar graphs with heights representing the probabilities for each outcome being chosen. I also chose for the *TinkerPlots* environment to include a dotplot to display the data. Other visual representation such as hat plots and boxplots can be created in the software along with visual representations of formal measures such as average, median, mode, and midrange. I made the choice not to discuss these formal measures in order to investigate what measures the students thought were important to calculate and discuss. These choices focused the study to use *TinkerPlots* to explore previously elicited models of sampling and allowed me to focus my research on using *TinkerPlots* to explore these models. I think that these choices were in line with the research questions that I aimed to investigate, but if I had allowed the students to create their own models it may have given me more insight to students’ developing models.

**Final Conclusions**

This study discusses how students’ reasoning of bootstrapping and informal inference could be developed in an introductory statistics course. The results demonstrated that it is possible for model development sequences to elicit and develop students’ models of bootstrapping. Key to eliciting these models of bootstrapping was students’ multiplicative
conception of the sample that allowed them to view the sample as a quasi-proportional representation of the population. *TinkerPlots* software allowed students to collect many samples and observe the growing distribution’s trend of approaching a bell curve shape. From this distribution, students interpreted dotplots to use informal measures to determine which outcomes were likely to occur versus unlikely. Although the methods developed by students were informal in nature, there were strong parallels to reasoning used in formal inference for hypothesis testing and the construction of confidence intervals.
Instructions

Answer each question as best as you can by checking the circle next to your answer. This test may contain material that you have not yet covered in your coursework. If there are questions that you do not know the answer or do not feel that you can make an educated guess, you may leave the question blank.
1) A research study randomly assigned participants into two groups. One group was given Vitamin E to take daily. The other group received only a placebo pill. The research study followed the participants for eight years. After the eight years, the proportion of each group that developed a particular type of cancer was compared.

What is the primary reason that the study used random assignment?

○ To ensure that a person does not know whether or not they are getting the placebo.

○ To ensure that the groups are similar in all respects except for the level of Vitamin

○ To ensure that the study participants are representative of the larger population.

2) A local television station in a city with a population of 500,000 recently conducted a poll where they invited viewers to call in and voice their support or opposition to a controversial referendum that was to be voted on in an upcoming election. Over 10,000 people responded, with 67% opposed to the referendum. The TV station announced that they are convinced that the referendum will be defeated in the election.

Select the answer below that indicates whether the TV station's announcement is valid or invalid, and why.

○ Valid, because the sample size is large enough to represent the population.

○ Valid, because 67% is far enough above 50% to predict a majority vote.

○ Invalid, because the sample is too small given the size of the population.

○ Invalid, because the sample may not be representative of the population.

3) Jean is considering two different routes for commuting to school. She sets up a randomized experiment where each day she tosses a coin to decide which route to take that day. She records the minutes of travel time for 5 days of travel on each route.

Route #1: 16, 11, 23, 7, 18 (Mean = 15, Standard Deviation = 6.20)
Route #2: 19, 15, 17, 16, 18 (Mean = 17, Standard Deviation = 1.58)

It is important to Jean to arrive on time for her classes, but she does not want to arrive too early. Based on the data gathered and Jean’s preferences, which route would you advise her to choose?

○ Route #1 since on average, she got to school quicker.

○ Route #2 since her travel times were more consistent.

○ Jean can choose either route since the times are about the same.
4) Indicate which distribution has the larger standard deviation.

○ A has a larger standard deviation than B
○ B has a larger standard deviation than A
○ Both distributions have the same standard deviation

5) One hundred student-athletes attended a summer camp to train for a particular track race. All 100 student-athletes followed the same training program in preparation for an end-of-camp race. Fifty of the student-athletes were randomly assigned to additionally participate in a weight-training program along with their normal training (the training group). The other 50 student-athletes did not participate in the additional weight-training program (the non-training group). At the end of the summer camp, all 100 student-athletes ran the same race and their individual times (in seconds) were recorded.

The mean speed of the training group was 44 seconds, and the mean speed of the non-training group was 66 seconds.

The standard deviation for the non-training group was 20 seconds. Consider the following possible values for the standard deviation of the training group. Which of these values would produce the strongest evidence of a difference between the two groups?

○ 10 seconds
○ 20 seconds
○ 30 seconds
Questions 6 and 7 ask you to think about factors that might affect the width of a confidence interval. For both questions, a confidence interval is shown as a horizontal line. The sample mean is represented by a solid dot in the middle of the confidence interval.

6) Imagine that two different random samples of test scores are drawn from a population of thousands of test scores. The first sample includes 250 test scores and the second sample includes 50 test scores. A 95% confidence interval for the population mean is constructed using each of the two samples.

Which set of confidence intervals (below) represents the two confidence intervals that would be constructed?
7) Imagine that a new random sample of 100 test scores is drawn from the thousands of test scores. A 99% confidence interval for the population mean and a 90% confidence interval for the population mean are constructed using this new sample.

Which set of confidence intervals (below) represents the two confidence intervals that would be constructed?
Imagine a candy company that manufactures a particular type of candy where 50% of the candies are red. The manufacturing process guarantees that candy pieces are randomly placed into bags. The candy company produces bags with 20 pieces of candy and bags with 100 pieces of candy. Which pair of distributions (below) most accurately represents the variability in the percentage of red candies in an individual bag that would be expected from many different bags of candy for the two different bag sizes?
Alicia was interested in whether offering people financial incentives can improve their performance playing video games. Alicia designed a study to test whether video game players are more likely to win a game when they receive a $5 incentive or when they simply receive verbal encouragement. Forty subjects were randomly assigned to one of two groups. The first group was told they would receive $5 if they won the game and the second group received verbal encouragement to “do your best” on the game. Alicia collected the following data from her study:

<table>
<thead>
<tr>
<th></th>
<th>$5 Incentive</th>
<th>Verbal Encouragement</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Win</td>
<td>16</td>
<td>8</td>
<td>24</td>
</tr>
<tr>
<td>Lose</td>
<td>4</td>
<td>12</td>
<td>16</td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td>20</td>
<td>40</td>
</tr>
</tbody>
</table>

Based on these data, it appears that the $5 incentive was more successful in improving performance than the verbal encouragement, because the observed difference in the proportion of players who won was

\[
\frac{16}{20} - \frac{8}{20} = \frac{8}{20} = 0.40
\]

In order to test whether this observed difference is only due to chance, Alicia does the following:

- She gets 40 index cards. On 24 she writes, "win" and on 16 she writes, "lose."
  - She then shuffles the cards and randomly places the cards into two stacks of 20 cards each. One stack represents the participants assigned to the $5 incentive group and the other represents the participants assigned to the verbal encouragement group.
  - She computes the difference in performance for these two hypothetical groups by subtracting the proportion of winning players in the "verbal encouragement" stack from the proportion of winning players from the "$5 incentive" stack.
  - She records the computed difference on a plot.
- Alicia repeats the previous three steps 100 times.
9) What is the explanation for the process Alicia followed?

○ This process allows her to see how different the two groups' performance would be if both types of incentive were equally effective.

○ This process allows her to determine the percentage of time the $5 incentive group would outperform the verbal encouragement group if the experiment were repeated many times.

○ This process allows her to determine how many times she needs to replicate the experiment for valid results.

10) Alicia simulated data under which of the following assumptions?

○ The $5 incentive is more effective than verbal encouragement for improving performance.

○ The $5 incentive and verbal encouragement are equally effective at improving performance.

○ Verbal encouragement is more effective than a $5 incentive for improving performance.
Below is a plot of the simulated differences in proportion of wins that Alicia generated from her 100 trials. Based on this plot, the one-sided p-value is 0.03.

11) Which of the following conclusions about the effectiveness of the $5 incentive is valid based on these simulation results?

○ The $5 incentive is not more effective than verbal encouragement because the distribution is centered at 0.

○ The $5 incentive is more effective than verbal encouragement because the distribution is centered at 0.

○ The $5 incentive is not more effective than verbal encouragement because the p-value is less than .05.

○ The $5 incentive is more effective than verbal encouragement because the p-value is less than .05.

For question 12, indicate whether the provided interpretation of the p-value is valid or invalid.

12) The p-value is the probability that the $5 incentive group would win more often than the verbal encouragement group.

○ Valid

○ Invalid
13) In Alicia’s experiment, there were 20 subjects randomly assigned to each group. Imagine a new study where 100 students were randomly assigned to each of the two groups. Assume that the observed difference in this new study was again 0.40 (i.e., that the proportion of wins for the $5 incentive group was 0.40 higher than the observed proportion of wins for the verbal encouragement group).

How would the p-value for this new study (100 per group) compare to the p-value for the original study (20 per group)?

○ It would be smaller than the original p-value

○ It would be the same as the original p-value

○ It would be larger than the original p-value
Instructor Debriefing Interview Protocol

- What did you think about some of the ideas that were tried?
- What did you think when … (something of interest from previous class occurred)?
- What went well? What didn’t? Why?
- What was supposed to happen? What did happen? Was there a difference between what we expected to happen and what did?
- If you were to do this lesson again, would you change anything?
Ophelia the Octopus

In the 2010 World Cup, Paul the Octopus (in a German aquarium) became famous for being correct in all eight of the predictions that he made. This included predicting Spain over Germany in a semifinal match. Before each game, two containers of food were lowered into the octopus's tank. The containers were identical except for the country flags of the opposing teams. The container that Paul opened was deemed his predicted winner.

Paul has since passed away, leaving an opening for sports predicting octopuses. Zookeepers at the [local] Zoo hope that their octopus Ophelia can fill this opening. The zookeepers' plan is for Ophelia to predict in a similar manner the winners of [local] University's next eight basketball games. The local newspaper has picked up on the story and contacted you to help them discuss Ophelia's predictions. The newspaper wants to discuss with you how they can interpret whatever results Ophelia comes up with.

Write a letter to the [local newspaper] with the range of outcomes that you think are likely for Ophelia to correctly predict. Include a description of the methods that you used to come to this conclusion. The newspaper staff can then use your method for stories about future sports predicting octopuses.

How can you use the cup of coins to help you write your letter to the [local newspaper]?
Paul the Octopus Controversy!

Name  

Zookeepers at the [local] Zoo came up the idea to have Ophelia predict sporting events because of the German octopus Paul. As discussed in the previous activity, Paul the Octopus became famous for being correct in all eight of the predictions that he made during the 2010 World Cup. Before each game, two containers of food were lowered into the octopus's tank. The containers were identical except for the country flags of the opposing teams. The container that Paul opened was deemed his predicted winner.

There was some controversy with Paul's predictions. Many people thought that there was no way that Paul could have guessed the winners of all eight matches. A popular soccer blog claimed in a post that Paul must be psychic! After writing your letter to the [local newspaper], what comment would you write on this blog post?
Ophelia the Octopus with *TinkerPlots*

Name ________________________________

Following in the footsteps of Paul the Octopus, the zookeepers at the [local] Zoo plan for their octopus Ophelia to predict the winners of [local] University's next eight basketball games. Before each game they will lower two containers of food into Ophelia's tank. The containers will be identical except for one being labeled with [local] University's mascot, and the other container with the opposing teams' mascot. The container that Ophelia opens will be deemed her predicted winner. This activity will use *TinkerPlots* software to simulate the possible outcomes for Ophelia's predictions.

Click on the file *Ophelia the Octopus.tp* to open the *TinkerPlots* software. On the screen you will see three windows. The sampler window contains a spinner, with half of the spinner labeled as 'Right' and the other half labeled as 'Wrong'. Click on the 'Run' button in the top left corner of the sampler window.

*TinkerPlots* just simulated Ophelia's predictions for the winners of eight basketball games. Each 'Right' response means that she choose correctly and each 'Wrong' response means that she chose incorrectly. Since half of the spinner is labeled 'Right' and the other half is labeled 'Wrong', Ophelia had an equal chance of correctly predicting each basketball game.

*TinkerPlots* stored Ophelia's predictions in the results window. The results window also shows the number of right predictions. The dotplot of results window contains a dotplot with Ophelia's number of correct predictions. Since only one set of eight games has been predicted, there is only one dot in the plot.

1) What was your result for Ophelia's eight predicted games?

2) Was this the result that you had expected? If not, what did you expect?
Appendix C

Continue simulating Ophelia's predictions by pressing the 'Run' button in the sampler window again. *TinkerPlots* again simulated Ophelia's predictions for eight games. The results window and results plot window now contain both of Ophelia's two predictions.

3) What was your second result for Ophelia's predicted games?

4) Was this the result that you had expected? If not, what did you expect?

Now create a total of 10 of Ophelia's predictions. Instead of hitting the 'Run' key 8 more times, find where is says 'Repeat +1' in the sampler window and double click on it. This will bring up the 'inspect sampler menu'. At the bottom of this menu it says 'Repeat 1 times'. Click on the 1, and enter 8. Close the 'inspect sampler menu' window and you will now see 'Repeat +8' in the sampler window. Next to the 'Run' button are the words 'Very Slow' with a sliding button underneath the words. This adjusts the speed of collecting samples. Slide the button to the right until it reads 'Medium'. Click 'Run'.

5) Sketch the dotplot in the result plot window below.
6) Are any of your predictions unusual? Why or why not?

7) Discuss with your group members if they have any unusual predictions? Why or why not?

8) How do your 10 predictions compare with your group members’ predictions?
9) Discuss with your group members how using TinkerPlots to simulate Ophelia's predictions compares with your methods using the coins in the first activity. Record your group's thoughts below.

10) If you were to write your letter to the [local newspaper] with this your new data, would your range of outcomes that you think are likely for Ophelia to correctly predict be the same or different? Why?

11) Do all members of your group think that the likely range of outcomes for Ophelia to correctly predict are the same, or are they different? Why?
Now create a total of 50 of Ophelia's predictions. Find where it says 'Repeat +8' in the sampler window and double click on it. In the bottom of the 'inspect sampler menu' it will say 'Repeat 8 times'. Click on the 8, and enter 40. Close the 'inspect sampler menu' window and you will now see 'Repeat +40' in the sampler window. Next to the 'Run' button use the sliding speed button to slide from 'Medium' to 'Medium Fast'. Click 'Run'.

12) Sketch the dotplot in the result plot window below.

13) How do your 50 predictions compare with your group members' predictions?

14) If you were to write your letter to the [local newspaper] with this new data, would your range of outcomes that you think are likely for Ophelia to correctly predict be the same or different? Why?
15) Do all members of your group think that the same likely range of outcomes for Ophelia to correctly predict are the same or are they different? Why?

Now create a total of 1000 of Ophelia's predictions. Find where it says 'Repeat +40' in the sampler window and double click on it. In the bottom of the 'inspect sampler menu' it will say 'Repeat 40 times'. Click on the 40, and enter 950. Close the 'inspact sampler menu' window and you will now see 'Repeat +950' in the sampler window. Next to the 'Run' button use the sliding speed button to slide from 'Medium Fast' to 'Fast + 10'. Click 'Run'.

16) Sketch the dotplot in the result plot window below.

17) How do your 1000 predictions compare with your group members' predictions?
18) If you were to write your letter to the [local newspaper] with this new data, would your range of outcomes that you think are likely for Ophelia to correctly predict be the same or different? Why?

19) Do all members of your group think that the likely range of outcomes for Ophelia to correctly predict are the same or are they different? Why?

20) Discuss with your group the similarities and differences between the simulations of Ophelia's predictions with the coins, using TinkerPlots for 10 predictions, 50 predictions, and using TinkerPlots for 1000 predictions.
21) Do you see advantages to using TinkerPlots instead of the coins?

22) Do you see disadvantages to using TinkerPlots instead of the coins?
Ophelia the Octopus - Take Two

Name ________________________________

In an earlier activity, you learned of the [local] Zoo's plan for their octopus Ophelia to predict the winners of [local] University's next eight basketball games. Using a cup of coins, you investigated the likely range of outcomes for Ophelia to correctly predict. You then wrote a letter to the [local newspaper] with your analysis.

Since writing that letter, you have used the software TinkerPlots to investigate Ophelia's predictions. Write another letter to the [local newspaper] with the range of outcomes that you think are likely for Ophelia to correctly predict. Use any of the data that you have gathered with the cup of coins or TinkerPlots in your analysis. Include a description of the methods that you used to come to this conclusion.
Nike Market Share

During the 2013 fiscal year Nike's sales revenue was over $25 billion. This made Nike the market leader in sport footwear and apparel. About 35% of all sneakers sold globally in 2013 were Nikes.

A marketing director from Nike has contacted you to help him determine if this global trend of Nike sneaker sales holds true for [local] University students. He wants to discuss sneaker sales for [local] University students at an upcoming meeting. The marketing director would like to know if it is reasonable to claim in his meeting that about 7 in 20 [local] University student sneaker purchases are Nikes.

Write a letter to the Nike marketing director. Discuss what claims about [local] University students' sneaker purchases you think that he should make. Explain to him how you came to your conclusions. He may be asked to justify these claims during his meeting.

You may use the bin of beads to investigate [local] University students' sneaker purchases. Each bead in the bin is a sneaker purchase by a [local] University student. Every clear bead is a Nike sneaker purchase.
Nike Market Share with *TinkerPlots*

Name ________________________________

About 35% of all sneakers sold globally in 2013 were Nikes. A marketing director from Nike contacted you to help him determine if the global trend of Nike sneaker sales held true for [local] University students. He wanted to know if it was reasonable to claim that about 7 in 20 [local] University student sneaker purchases are Nikes.

You used a bin of beads to investigate [local] University students' sneaker purchases. Each bead in the bin was a sneaker purchase by a [local] University student. You then wrote a letter to the Nike marketing director that discussed what claims about [local] University students' sneaker purchases you thought that he should make. The letter also explained how you came to your conclusions.

Click on the file *Nike Market Share.tp* to open the *TinkerPlots* software.

Just as the last activity with *TinkerPlots*, there are three windows. The sampler window contains a mixer with the same number of balls as there were beads in the bin used in class. The balls are labeled 'N' for Nike sneaker purchases and 'A' or 'O' for other brands of sneaker purchases. When the sampler is run, the results window keeps track of the number of Nike sneaker purchases in each sample. The results plot window shows a dotplot of the number of Nike sneaker purchases in each sample.

Click 'Run' in the sampler widow and watch TinkerPlots collect one sample of 20 sneaker purchases of [local] University students. Using the spatula and bin of beads, you chose 20 beads at once. The sampler chooses one ball at a time (without replacing that ball) until 20 balls are chosen.

The class collected a total of 120 samples from the bins of beads. Below is the dotplot of sneaker purchases when all groups' samples were combined.
Simulate a total of 120 samples with *TinkerPlots*. Double click on 'Repeat +1' in the sampler window. In the 'Inspect Sampler' window, change the setting from 'Repeat 1 Times' to 'Repeat 119 Times'. Close the 'Inspect Sampler' window. Adjust the speed of taking samples from 'Very Slow' to 'Fast +10'. Click 'Run' in the sampler window.

1) Sketch the dotplot below. At the top of the dotplot are the number of dots in each column of the plot.

2) How does this dotplot compare to the dotplot of 120 samples that was created in class? Be specific about any similar or different features in the two dotplots.
3) In the previous activity, you used data to write a letter to the Nike marketing director. The letter discussed what claims about [local] University students' sneaker purchases you thought that he should make. Using the dotplot created in *TinkerPlots*, what range of Nike sneaker purchases do you think are likely for 20 [local] University student purchases? Be specific about what features of the dotplot made you choose your range of values.

4) Is there a certain percentage of the data points that you believe should be in the likely range of the number of Nikes for 20 [local] University students' sneaker purchases? What percentage and why?

5) Determine the percentage of the data points in the *TinkerPlots* dotplot that are within your likely range of the number of Nikes for 20 [local] University students' sneaker purchases. The numbers at the top of the dotplot are the number of dots in each column.
6) Discuss with your group members the percentage of data points that are in each of your likely ranges of the number of Nikes for 20 [local] University students’ sneaker purchases. Are they similar or different? Why?

7) Do these percentages influence your decision for the likely range of the number of Nikes for 20 [local] University students’ sneaker purchases? Why or why not?

8) An advantage of TinkerPlots over the bin of beads is that many samples can be simulated very quickly. How many samples should you simulate to be confident in your claims? Why? **Do not run another simulation yet.**
9) If you were to simulate the number of samples that you answered in the previous question, what would the dotplot of the results look like? Do not run another simulation yet. Sketch what you think the plot will look like below. Describe the differences and similarities to your dotplot with 120 samples.

10) Using TinkerPlots, simulate the number of samples that you answered in question 8. Sketch the dotplot below. Describe the differences and similarities to what you through the dotplot would look like and your simulated dotplot.
11) Using the dotplot from this new simulation, what range of Nike sneaker purchases do you think are likely for 20 [local] University student purchases? Be specific about what features of the dotplot made you choose your range of values.

12) Is this range from your new simulation similar or different from the range that you predicted from the dotplot of 120 samples? Why?

13) Using the dotplot from this new simulation, determine the percentage of the values that are within your likely range of the number of Nikes for 20 [local] University students' sneaker purchases.
14) Compare the percentages of data points in your likely ranges of Nike sneaker purchases when using 120 samples and the number of samples that you answered in question 8. Discuss with your group members whether these percentages are similar or different and why.

15) What if you didn't have the sampler in *TinkerPlots* or the bin of beads used in class? What exactly does each bead represent? Come up with a plan with your group members of how you would collect the data needed to make claims about [local] University student sneaker purchases.
John Slusher  
One Bowerman Drive  
Beaverton, OR 97005  
October 19, 2014

[Instructor]’s Statistics Students  
[School address]

Dear [Instructor]’s Statistics Students:

Thank you for your analysis of the sneaker purchases of [local] University Students. It was very helpful in my recent meeting and made us consider the sales figures for competing brands of sneakers.

About 35% of all sneaker purchases globally are Nikes. Adidas sneakers are the second most popular brand. About 20% of all sneaker purchases globally are Adidas. I am interested to know if this difference of about 15% between Nike and Adidas sneaker purchases is true for sneaker purchases of [local] University students.

In your letter, you described using a bin of beads to analyze sneaker purchases. Each bead in the bin was a sneaker purchase by a [local] University student. The clear beads were Nike sneaker purchases. If the blue beads are Adidas sneaker purchases, can you use the bin of beads to find the likely range for the difference in percentage between Nike and Adidas sneaker purchases for [local] University students?

Sincerely,

John Slusher  
Executive Vice President of Global Sports Marketing
Wegmans carries many types of nuts, dried fruits and candies in their bulk food section. The manager of bulk food is always interested in bringing new types of food for her customers to try. She recently ordered a sample of a new brand's mixed nuts. From past experiences, she has determined that customers prefer mixed nuts with fewer peanuts. She plans to order a large shipment of mixed nuts and is considering this new brand. Before she orders, the manager wants to know more information about the percentage of peanuts in this new brand.

From this one sample of mixed nuts, the manager has asked that you determine a likely range for the percentage of peanuts in the entire brand of mixed nuts. She would also like to know the methods that you develop to come to your conclusion. Your methods may of use for future bulk food purchases.

The bag of sticks is the sample of mixed nuts. Sticks marked with a 'P' are peanuts. Those not marked are other types of nuts.
Trident Gum

Below is a popular marketing slogan from Trident gum.

If you were to ask 5 dentists if they would recommend Trident for their patients who chew gum, what is the likely range of dentists that would say that they do?

How could you use 5 note cards to investigate this likely range of dentists?
Mixed Nuts with *TinkerPlots*

Name ________________________________

The manager of bulk food at Wegmans ordered a sample of mixed nuts. The sample contained 7 peanuts out of a total of 25 nuts. From this one sample of mixed nuts, the manager asked you to determine a likely range for the percentage of peanuts in the entire brand of mixed nuts.

1) Using a bag of 25 Popsicle sticks, with 7 marked as 'p' for peanut, how did you determine the likely range for the percentage of peanuts in the entire brand of mixed nuts?

2) How are the methods that you used to come to your conclusions about the mixed nuts similar to the methods used to determine the likely range of Nike sneaker purchases by [local] University students?
Open the file *Mixed Nuts.tp*. Just as in the previous *TinkerPlots* activities, there is a sampler window to generate samples, a results window to store the samples, and a results plot to display the samples graphically.

3) How many samples should you generate to predict the likely range of the percentage of peanuts in this brand of mixed nuts? Is there a way to know if you've generated enough samples?

4) Generate the number of samples that you stated in the previous question. Sketch a graph of the dotplot on the first axes on page 7. Next to the dotplot, write the number of peanuts in your original sample. From this one sample of mixed nuts, what do you predict to be the likely range for the percentage of peanuts in the entire brand of mixed nuts? Illustrate this predicted range on your dotplot.

5) Why did you choose your range of values in the previous question?
The bulk food manager at another Wegmans location also received a sample of this brand's mixed nuts. Each person in your group should collect a new sample of 25 mixed nuts from the bag in the front of the room.

6) How many peanuts were in your new sample of 25 mixed nuts?

7) If this one new sample was the only sample available, what would you estimate to be the likely range for the percentage of peanuts in the entire brand of mixed nuts? Do not take into account your previous sample or group members’ samples. Why?

Next, simulate new samples of mixed nuts from the sample that you just gathered. First, you'll need to erase the old data. Click on the options button in the upper right corner of the results window. The results window is the bottom left window and collects the data from all of the samples. Then click on 'Delete all Results Cases'. This will remove the data from your previous simulation. Next, you will need to adjust the sampler for your new sample of mixed nuts. In the bottom left corner of the sampler window, click on the icon of a lock. Now if you click on any of the balls in the sampler, you are able to rename them. Rename the balls so that you have enough lowercase 'p's and 'o's for your new sample of mixed nuts.

8) Simulate sample data from this new sample and sketch a graph of the dotplot on the second axes on page 7. Next to the dotplot, write the number of peanuts in your new sample. From this new sample of mixed nuts, what do you predict to be the likely range for the percentage of peanuts in the entire brand of mixed nuts? Illustrate this predicted range on your dotplot.
9) How are the two dotplots simulated from two different samples similar? How are they different?

10) Is there a way to determine which prediction for the likely range of percentages of peanuts is better?

When the manager of bulk food at Wegmans first ordered the sample of mixed nuts, she also asked the company that produced the nuts what percentage of their brand was peanuts. Weeks after she initially asked, they responded that the percentage varies slightly for each batch that they produce, but the current batch contains approximately 21% peanuts.

11) Was 21% peanuts in both of your predicted ranges for the likely percentage of peanuts? What does that tell you?
12) Was 21% peanuts in all of the class' predicted ranges for the likely percentage of peanuts? What does that tell you?

13) If the percentage of peanuts in all of the brand's mixed nuts is 21% how could you simulate the likely percentage of peanuts in samples of 25 nuts?

14) Open the file *Mixed Nuts_21.tp*. How does this file simulate the likely percentage of peanuts in samples of 25 mixed nuts?
15) Now that you know the percentage of peanuts in the entire brand's mixed nuts, simulate sample data and sketch a graph of the dotplot on the third axes on page 7. What do you predict to be the likely range for the percentage of peanuts in the one sample of 25 mixed nuts? Illustrate this predicted range on your dotplot.

16) How are the three dotplots similar? Why? How are they different? Why?

17) Were all of your class' samples of mixed nuts in your predicted likely range for the percentage of peanuts in the one sample of 25 mixed nuts? How does this relate to the predictions made from a single sample?
Dotplots
Comparing Brands of Mixed Nuts

The manager of bulk food at Wegmans decided to order another sample of mixed nuts, but from a different brand. She wants to know which brand has the smaller percentage of peanuts. From these two samples of mixed nuts, she wants to find the likely range of the differences in the percentages of peanuts for the two brands.

How can you convince her of which brand to order?

The two bags of sticks, marked A and B, are the samples of mixed nuts. Sticks marked with a 'P' are peanuts. Those not marked are other types of nuts.
References


knowledge of mean, median, and mode. *Mathematical Thinking and Learning, 8*(1), 37-63.


(www.minitab.com)


NAME OF AUTHOR: Jeffrey Allen McLean

PLACE OF BIRTH: Rochester, New York

DATE OF BIRTH: September 10, 1979

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

   Syracuse University, Syracuse, New York
   University of Connecticut, Storrs, Connecticut
   Rochester Institute of Technology, Rochester, New York

DEGREES AWARDED:

   Master of Science in Mathematics, 2008, University of Connecticut
   Bachelor of Science in Applied Mathematics, 2005, Rochester Institute of Technology

PROFESSIONAL EXPERIENCE:

   Visiting Assistant Professor, Sonoma State University, 2015-Present