Statistical reasoning with the sampling distribution

Bridgette Jacob
Onondaga Community College, Syracuse University, New York
Helen M. Doerr
Syracuse University, New York

Introduction

Formal statistical inference consists of estimating population parameters with confidence intervals and testing conjectures about population parameters using hypothesis tests. The study of formal statistical inference in an introductory statistics course generally follows the study of descriptive statistics, probability, and the sampling distribution. The difficulties students have in bringing these concepts together to conduct formal statistical inference are well known in the field of statistics education (Castro Sotos, Vanhoof, Noortgate, & Onghena, 2009; Garfield & Ben-Zvi, 2008; Konold, 1989; Konold *et al.*, 2011; Tversky & Kahneman, 1974). Recently, "informal" statistical inference has been studied as a means to bridge the gap between descriptive statistics and formal statistical inference. In informal statistical inference, students draw conclusions about populations based on data without the formalities of constructing a confidence interval or conducting a hypothesis test. Instead, the data are examined to understand the main features (*e.g.*, center and spread) to determine what evidence this data might provide about the population.

An understanding of how informal inferential reasoning develops may help to build the bridge to formal inferential reasoning, the reasoning that ties together the concepts of descriptive statistics, probability, and the sampling distribution. Recent research efforts in statistics education have focused on informal statistical inference to understand how students begin to reason about data (Ben-Zvi, 2004; Pfannkuch, 2006; Pratt, Johnston-Wilder, Ainley, & Mason, 2008). Makar and Rubin (2009) defined informal inferential reasoning as generalizing about a population using sample data as evidence while recognizing the uncertainty that exists. While researchers are building definitions of informal inferential reasoning and frameworks for researching its development (Makar & Rubin, 2009; Pfannkuch, 2006; Zieffler, Garfield, DelMas, & Reading, 2008), exactly how informal inferential reasoning develops and how students demonstrate such reasoning across a range of contexts is still under investigation. The research reported here was part of a larger study designed to add to the understanding of that development by investigating how

secondary students demonstrated informal inferential reasoning throughout a year-long course in introductory statistics in the United States. The research questions investigated in this study were: 1) How do students reason informally with the sampling distribution; and 2) What is the relationship between students' informal inferential reasoning with the sampling distribution, their prior informal inferential reasonings and their subsequent formal inferential reasoning?

Background

Formal statistical inference involves conducting a hypothesis test or constructing a confidence interval with data from a sample, then drawing appropriate conclusions about a population. Although students can be taught the procedures of formal statistical inference, many of which they will remember and demonstrate, this does not necessarily mean they are able to draw and interpret appropriate conclusions about populations based on data or fully comprehend the assumptions behind and implications of those conclusions. An understanding of how the underlying concepts work together is required to make decisions based upon an appropriate analysis and interpretation of the data.

The Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report (2010) was developed by statisticians and statistics educators and endorsed by the American Statistical Association (ASA) in the United States. The report emphasizes conceptual understanding of the procedures necessary for statistical analysis. Among the goals for inferential reasoning put forth in the report are the following:

Students should understand the basic ideas of statistical inference, including:

- the concept of a sampling distribution and how it applies to making statistical inferences based on samples of data (including the idea of standard error);
- the concept of statistical significance, including significance levels and *p*-values; and
- the concept of confidence interval, including the interpretation of confidence level and margin of error. (ASA, 2010, p. 12)

Research in statistics education has shown that understanding the underlying concepts taught throughout an introductory statistics course often does not provide students with the support to achieve the conceptual understanding necessary for formal statistical inference. What may appear to be the next step in learning statistics to those who are seasoned in the subject is actually a large chasm for many introductory statistics students.

A conceptual understanding of the sampling distribution is necessary for formal statistical inference. Saldanha and Thompson (2002) designed a study to examine students' developing ideas about repeated sampling and the sampling distribution while they participated in instruction on these topics. Their study was based on prior research that found students focused on the statistics of samples, the sample mean, for example, rather than on how these statistics were distributed. Therefore, the instruction in their study

stressed two themes: "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). These researchers found that the majority of the students still compared a single sample statistic to the population parameter rather than to the sampling distribution of all such statistics when asked to determine if the sample statistic was unusual. Saldanha and Thompson claim that a conception of sample in which the distinctions among the population, the individual samples taken from the population, and the distribution of many such samples are made will help students to understand why statisticians can be confident in inferring about the population based on data from a single sample.

In a research project spanning seven years, Chance, delMas, and Garfield (2004) used interactive software designed to assist students in making these distinctions among the population, samples, and the sampling distribution. Some of the common misconceptions held by the students in their studies were: the belief that the sampling distribution should look like the population; predicting that sampling distributions for small and large sample sizes have the same variability; the belief that sampling distributions for large samples have more variability; and a lack of understanding that a sampling distribution is a distribution of sample statistics (p. 302).

These studies together demonstrate the difficulties students encounter when introduced to the sampling distribution. Many of the difficulties surround students' misunderstandings of the effects of sample size on the variability of the sampling distribution. In addition, once students have an initial perception of the sampling distribution, placing the sampling distribution in relation to the population and individual samples remains problematic.

Theoretical frameworks

Two theoretical frameworks were used for this study: one that influenced the sequence of informal statistical inference tasks used with the introductory statistics students; and a second that assisted with the evaluation of students' informal inferential reasoning as they completed these tasks.

The framework for the sequence of tasks used to explore students' informal inferential reasoning was a modified version of the task design framework developed by Zieffler, Garfield, delMas, and Reading (2008). These authors proposed three categories of tasks to conduct research on informal inferential reasoning:

- estimate and draw a graph of a population based on a sample;
- compare two or more samples of data to infer whether there is a real difference between the populations from which they were sampled; and
- judge which of two competing models or statements is more likely to be true. (p. 47)

We modified the first category into a task in which students estimated a population parameter based on their own random sampling. The first two task categories were then interchanged to provide a clear link to students' informal inferential reasoning as it developed based on the curriculum and sequencing of topics in their introductory statistics course. This laid the foundation for the three informal statistical inference tasks used in this study.

Therefore, the first task administered in this study corresponded to the second category in the Zieffler *et al.* (2008) framework and had students comparing distributions of data to make an informal inference. The design of this task drew on the research of difficulties students encounter with the concepts of variation and distribution in descriptive statistics (Bakker & Gravemeijer, 2004; Ben-Zvi, 2004; Kelly & Watson, 2002; Makar & Confrey, 2005; Reading & Shaughnessy, 2000; Shaughnessy, Canada, & Ciancetta, 2003; Watson, 2002; Watson & Moritz, 1999). Ben-Zvi (2004) found that comparing distributions would help students progress from a local perspective, within a data set, to a global perspective of describing variability between data sets. Watson and Moritz (1999) concluded that students who were able to see several aspects of data sets working together as a whole were best poised to make inferences when comparing those data sets. Their investigations included data sets of the same size and data sets of different size requiring a proportional understanding of variation and distribution. Watson and Moritz conjectured that comparing distributions of data provided students with the opportunity to gain a deeper understanding of the basic concepts of descriptive statistics which might foster the development of their informal inferential reasoning.

The second task was related to the first category in the Zieffler *et al.* (2008) framework. Working on this task, students made inferences about unknown population probabilities based on their empirical sampling. Research on students' misconceptions of sampling and basic probability (Konold, 1989; Konold *et al.*, 2011; Tversky and Kahneman, 1974) was the basis for this second task modeled after the Bone Problem task used by Konold *et al.* (2011). These researchers theorized that giving students the opportunity to estimate the probability of an event that could not be summarized with a theoretical probability (unlike the probability of obtaining a sum of seven when tossing two die) supports their informal inferential reasoning by providing a conceptual understanding of the uncertainty that exists when drawing an inference and a level of confidence in their inferences. Estimating in this manner also provides students with the opportunity to consider the importance of random samples and large samples. In their research, Konold *et al.* found students lacked an understanding of these concepts.

The third task was related to the third category proposed by Zieffler *et al.* (2008) and culminated with students making an informal inference based on a sample of their data and the corresponding sampling distribution. The parts of this task stemmed from research of students' difficulties in understanding the sampling distribution (Chance, delMas, & Garfield, 2004; Saldanha & Thompson, 2002) and were designed to support the students in making the distinction between the population distribution, the distribution of a single sample taken from the population, and the distribution of the sample

statistics of many samples. Making an informal inference by examining where a sample statistic is situated in comparison to all such samples in the sampling distribution may help in developing students' informal inferential reasoning and is necessary for formal statistical inference.

This sequence of three tasks provided a means to gain insight into students' informal inferential reasoning throughout their study of introductory statistics. To assist in the analysis of students' informal inferential reasoning as they worked through the series of tasks, the three principles essential to informal statistical inference developed by Makar and Rubin (2009) were used:

> (1) *generalization*, including predictions, parameter estimates, and conclusions, that *extend beyond describing the given data*; (2) the use of data *as evidence* for those generalizations; and (3) employment of probabilistic language in describing the generalization, including informal reference to levels of certainty about the conclusions drawn. (p. 85)

Evidence of students' informal inferential reasoning was determined by the extent to which they (1) made an inference based on the data, (2) used the data as evidence for their inference, and (3) used probabilistic language to indicate a level of certainty in their inference.

Design and methodology

This study was designed to follow the development of introductory statistics students' informal inferential reasoning leading to their formal inferential reasoning. A series of four task-based interviews (Goldin, 2000) with student pairs were conducted to examine this development. The first three task-based interviews included informal statistical inference tasks that took place in the following order in conjunction with the progression of the class curriculum: (1) a comparing distributions task; (2) a sampling and probability task; and (3) a sampling distribution task. These tasks were designed to engage students in informal inferential reasoning as they progressed through the curriculum of their introductory statistics courses. To complete the study, the fourth task-based interview contained formal statistical inference tasks and took place near the end of the course. This sequence of task-based interviews provided a means to gain insight into students' informal inferential reasoning throughout their study of introductory statistics and their formal inferential reasoning. In this paper, we mainly focus on the students' work during the sampling distribution task when students were asked to draw an informal conclusion by taking a random sample and comparing it to the related sampling distribution. Because we were interested in the development of students' informal inferential reasoning, we examined how students' responses drew on their earlier reasoning in the comparing distributions and the sampling and probability tasks. We then examined how their reasoning in the sampling distribution task impacted their responses in the formal statistical inference task.

The task-based interviews (Goldin, 2000) used explicit interview protocols allowing students to think about their responses without critiquing for correctness and included tasks with appropriate content for students to grasp. The interviews were structured on key statistical concepts that gave students a variety of ways to demonstrate their understanding, and involved students in free problem solving while they interacted with another student. The interview tasks were designed with multiple parts, increasing in complexity.

The students taking part in this study were enrolled in introductory statistics courses for college credit in their secondary schools. They were 16 to 18 years of age and had successfully completed at least two mathematics courses, including a required algebra course. These students came from one of eight statistics classes taught by four different high school mathematics teachers from two high schools. There were four such classes at each high school with each teacher teaching two classes. These statistics classes met for approximately three and one-half hours each week for the 40-week school year.

The student pairs participating in the study were selected with the assistance of their classroom teachers. The teachers identified students with a range of prior achievements in mathematics, who had good attendance records, were communicative, and would work well together. These criteria provided assurance that students would have taken part in the classroom learning and would express their understandings of the key statistical concepts and their informal inferential reasoning. Seven pairs of students completed all four tasks in the study.

The Comparing Distributions Task

The first task involved students in informal inferential reasoning by asking them to make an inference about two populations based on sample data. This task followed the students' classroom study of descriptive statistics including graphical displays of data (*e.g.* histograms and boxplots), measures of center (mean and median), and measures of variability (range and standard deviation).

There were five parts to this task, each comparing data of children's test scores from two classes. These comparisons were modified from questions used by Watson and Moritz (1999). The student pairs were asked if the classes scored equally well or if one of the classes scored better. The first part required a comparison of the measures of center with one of the classes clearly scoring better. The second part also required a comparison of measures of center, however, the students had to take the shape (one was skewed right and one was skewed left) into consideration. In the third part, students were shown two distributions with the same mean but different variability. Adding a layer of complexity, the fourth part showed two classes of different size in which students had to reason proportionally in determining which class performed better. To complete the fifth and final comparison, students needed to combine their proportional reasoning with the concept of variability. In addition to the Makar and Rubin (2009) framework used to analyze students' informal inferential reasoning, student responses were analyzed based on key statistical concepts to determine if they: (1) used means to compare distributions;

(2) used variation to compare distributions; (3) recognized the effects of skewness on the mean; and (4) reasoned proportionally when distributions were of different size.

The Sampling and Probability Task

During this task students estimated an unknown probability by collecting their own data. This followed their study of random sampling (sampling methods generating samples representative of the population that avoid bias), the Law of Large Numbers (as the number of independent repeated trials increases, the relative frequency approaches the probability of the event), and basic probability rules (*e.g.* the probabilities in a model sum to one and the probabilities of complements).

Modeled after the Bone Problem used in the Konold *et al.* study (2011), the task began by asking students to estimate the probability that a Monopoly house would land upright when it was tossed. The students collected data by tossing the houses and estimated this probability. The results from 1,000 tosses of the Monopoly houses were then revealed and students were asked if this helped in estimating this probability. This was done to determine the extent of their understanding of the long-run characteristic of probability. For the second part of the task, students were shown a container of multi-colored beads and asked to estimate the proportion of green beads. For sampling, the students had a device that drew samples of 32 beads at a time. Student responses were analyzed based on key statistical concepts to determine if they: (1) took random samples; and (2) used the long-run characteristic of probability. This was in addition to the Makar and Rubin (2009) framework used to analyze students' informal inferential reasoning,

The Sampling Distribution Task

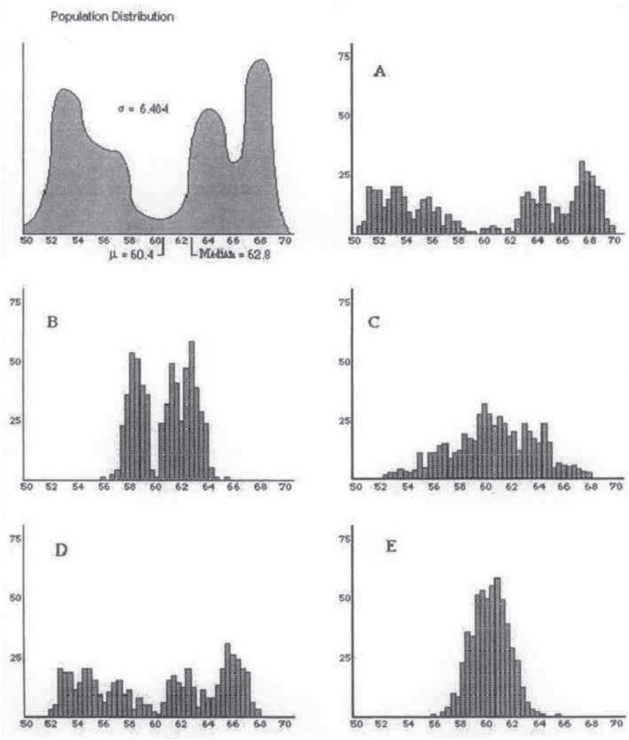
This task culminated with students making informal inferences about a population based on how a sample of data compared to a large number of such samples of data. This followed students' classroom study of sampling distributions during which they had been exposed to the normality of sampling distributions and to the effects of sample size on the variability of a sampling distribution (the larger the sample size, the less the variability in the sampling distribution).

This task began with part of an activity developed and used by Chance, delMas, and Garfield (2004) in which students were shown a tri-modal distribution (Figure 1). Students predicted what the sampling distribution would look like by choosing from five graphs of distributions and then answered questions about the effect of sample size on the variability of the distributions. Students were asked to explain how they chose the graphs and what role variability played in their choices.

Sampling Distribution Task

PART 1

The distribution for a population of test scores is displayed below on the left. Each of the other five graphs labeled A through E represents possible distributions of sample means for random samples drawn from the population.



1. Which graph represents a distribution of sample means for 500 samples of size 4? (circle one) A B C D E

2. I expect this sampling distribution to have (circle one) less, the same, more variability than the population?

3. Which graph represents a distribution of sample means for 500 samples of size 16? (circle one) A B C D E

4. I expect this sampling distribution to have (circle one) less, the same, more variability than the first sampling distribution?

Figure 1 — Part one of the sampling distribution task.

In the second part of the task, students viewed the Random Rectangle simulation in *Fathom*, shown in Figure 2, which was designed to help them understand distinctions among the population distribution, the distribution of a single sample, and the distribution of the sample means. The population of rectangles, labelled with their corresponding areas, is shown on the left and the graph of the areas of the total population of rectangles is in the upper middle. The Sample of Rectangles graph (in the lower middle) displays the distribution of a single random sample of 10 rectangles. The Measures from Samples of Rectangles graph in the lower right displays the sampling distribution of the mean areas. Students were able to watch a demonstration that animated how each sample was taken from the population, graphed, and then the mean area from each sample was added to build the sampling distribution.

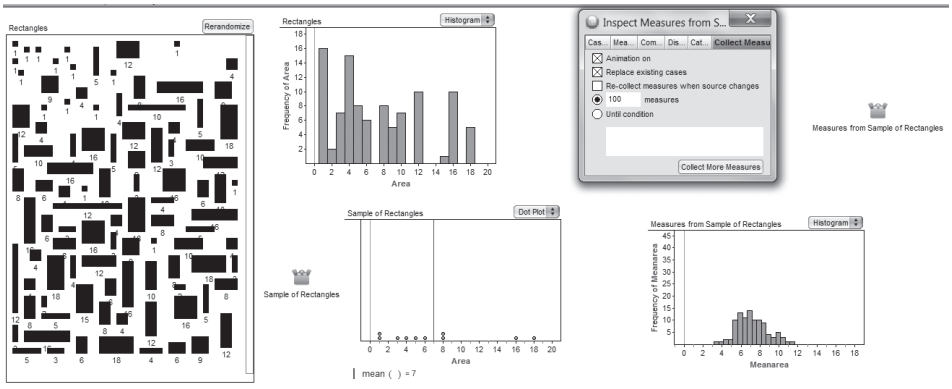


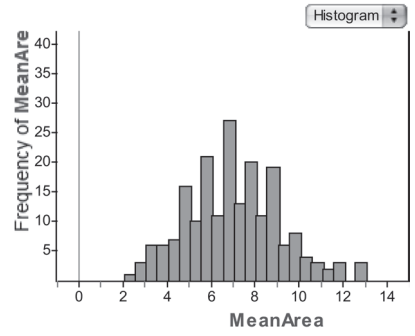
Figure 2 — Screen shot of Random Rectangle simulation.

The students were then presented with three sampling distributions generated from the Random Rectangles simulation activity in *Fathom* (Figure 3). The sample size increased from five to 10 and then to 25 as the distributions of mean areas were graphed. Students were asked which average areas would be likely and which would be rare or unlikely based on each of the sampling distributions.

The third part of the sampling distribution task was influenced by the work of Saldanha and Thompson (2002) who found that even after instruction on the sampling distribution, students tended to compare the results from a sample to the distribution of the original population rather than to the sampling distribution. Therefore, the task concluded by returning to the Monopoly houses used in the sampling and probability task to test the hypothesis that a Monopoly hotel had the same probability of landing upright as a house. Students were shown a sampling distribution of sample proportions of houses landing upright generated from 200 samples of 10 houses to assist them in making this informal inference (Figure 4). In addition to the Makar and Rubin (2009) framework used to analyze students' informal inferential reasoning, students' responses

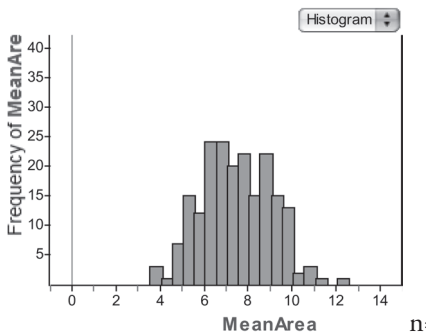
were analyzed to determine if they: (1) distinguished between the sampling distribution (approximately normal distribution of a statistic) and the population distribution; and (2) recognized that the variability of the sampling distribution was less than that of the population and decreased as the sample size increased.

PART 2



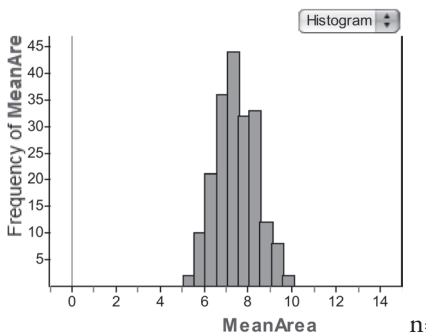
(1) Approximately what values of the sample mean for samples of size 5 would be reasonably likely?

(2) Rare events are defined as those that will occur less than 5% of the time. What values of the sample mean for samples of size 5 would you consider rare?



(1) Approximately what values of the sample mean for samples of size 10 would be reasonably likely?

(2) What values of the sample mean for samples of size 10 would you consider rare?



(1) Approximately what values of the sample mean for samples of size 25 would be reasonably likely?

(2) What values of the sample mean for samples of size 25 would you consider rare?

Figure 3 — Part two of the sampling distribution task.

PART 3

Below is the distribution for 200 samples of size 10 for the proportion of Monopoly houses that landed upright.

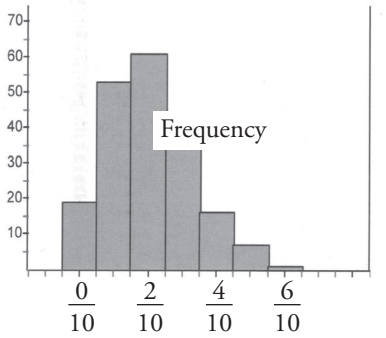


Figure 4 — Sampling distribution for 200 tosses of 10 Monopoly houses.

The Formal Statistical Inference Task

The final task encompassed formal statistical inference by asking students to interpret confidence interval estimates of population parameters and interpret the results of a hypothesis test. This followed classroom instruction on these topics. For this task, the same container of multi-colored beads used in the sampling and probability task (when they were asked to estimate the proportion of green beads) was used. The task began with students viewing 10 confidence intervals for the proportion of red beads in the container which were constructed from random samples. Students were asked what these confidence intervals revealed about the proportion of red beads in the container. The students were then asked to construct a confidence interval for the proportion of red beads in the container by drawing a sample with the device that captured 32 beads. The task concluded with students conducting a hypothesis test using their sample of red beads to determine if they agreed or disagreed with a conjecture made about the proportion of red beads in the container.

Data Collection and Analysis

The task-based interviews with the seven pairs of students were videotaped and transcribed. The transcriptions and videotapes of the interviews were analyzed in three phases. In the first phase, the transcripts were coded for the key statistical concepts of each task and the three main principles of informal statistical inference described by Makar and Rubin (2009). In the second phase of analysis, the transcripts and videotapes were examined by task and coded by the similar/different inferences drawn by the

students and the similar/different methods they used and reasonings they gave for their inferences. In the final phase of analysis, these inferences, methods and reasonings were grouped to determine how they were related across the four tasks.

Results

The results reported here include students' responses during the sampling distribution task and the formal statistical inference task. We found their responses to the sampling distribution task had some grounding in their responses during the comparing distributions and sampling and probability tasks and impacted their responses in the formal statistical inference task.

During the sampling distribution task, students demonstrated an understanding of the characteristics of the sampling distribution and could identify likely and unlikely sample proportions in relation to a sampling distribution. However, when asked to draw a conclusion based on a sampling distribution, students were not confident that they could draw a conclusion based on a single sample and wanted to take several samples before drawing a conclusion. During the formal statistical inference task, students also exhibited the propensity to take several samples rather than a single sample to form a confidence interval and demonstrated procedural rather than conceptual knowledge while completing the formal statistical inference task.

The Sampling Distribution Task

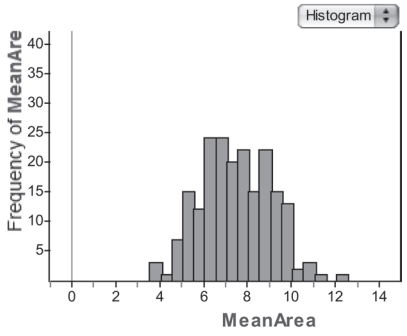
We found that most students had a base knowledge of the sampling distribution and its characteristics. In the first part of the sampling distribution task when students were shown the tri-modal population distribution (Figure 1), all seven pairs of students chose sampling distribution graphs that were approximately normal in shape. Six of them also correctly identified the effect of sample size on the variability of the sampling distribution. Only one pair incorrectly identified the variability of the sampling distribution for a sample of size four; however, they did correctly identify the variability as less for the sample size of 16.

We also found during the sampling distribution task that most students demonstrated an understanding of the probabilities and variability associated with the normality of the sampling distributions of mean areas of random rectangles. Following the *Fathom* demonstration (Figure 2), the students were shown the three sampling distributions of mean areas generated with the simulation for 100 samples of sizes five, 10, and 25 rectangles (Figure 3). When asked how the three distributions compared to one another, six of the pairs of students referred to these distributions as becoming more centered or having the same mean. Five of the pairs also referred to the decrease in variability as the sample size increased, as did this student:

Interviewer: So we went from a sample size of 5, then to 10, now to 25. So how about this one [of sample size 25]?

Jared: This one's even more compact. The last one [of sample size 10] got all the way out to like 12. This one hasn't gone past 10 [referring to maximum mean area].

The remaining pair referred to the decrease in variability alone, mentioning the formula $(\sigma/\sqrt{n})$ for standard error in support of this decrease. Students were then asked what mean areas would be likely and which would be rare or unlikely for each of the sampling distributions. Figure 5 illustrates a typical response from the student pairs for the samples of size 10 and 25. The students identified ranges of outcomes surrounding the peak of the approximately normal distributions as likely and those in the tails as rare. In doing so, they demonstrated proper probabilistic reasoning related to these sampling distributions.



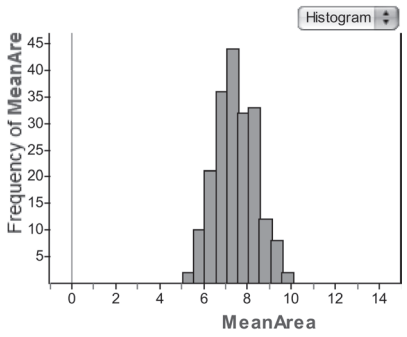
n=10

(1) Approximately what values of the sample mean fo samples of size 10 would be reasonably likely?

5-10

(2) What values of the sample mean for samples of size 10 would you consider rare?

10 and higher and 4 and below



n=25

(1) Approximately what values of the sample mean fo samples of size 25 would be reasonably likely?

6-9

(2) What values of the sample mean for samples of size 25 would you consider rare?

5 and less and 9 and higher

Figure 5 — Example of student work in part two of the sampling distribution task.

In an effort to bring the characteristics of the sampling distribution and the probabilities related to the sampling distribution together to make an informal inference, the students were shown the sampling distribution in Figure 4. The interviewer explained to the students that this sampling distribution was generated by tossing 10 Monopoly houses 200 times, recording the proportion of houses landing upright. The students were then asked if they could determine whether the probability that a Monopoly hotel

would land upright was the same as that for a house. The hotels were slightly larger with a rectangular rather than a square base like the houses. Both the houses and hotels were available for students to manipulate.

The majority of students believed that to accurately infer about the hotels, they would need to compare distributions and preferably, distributions of the same size. Four of the seven pairs expressed that they would need to generate a sampling distribution in the same manner as that shown for the houses. This occurred even though these students had demonstrated an understanding of the characteristics of the sampling distribution and had identified outcomes that would be considered likely and rare in relation to sampling distributions in the first two parts of this task. These pairs thought that they would need to toss 10 hotels 200 times to make a fair comparison, as expressed by Jared.

> Interviewer: So is there a way that you could make some type of determination so that you could tell me that you think the probability [of the hotel] is the same [as the house] or that you think it's different.
>
> Jared: Well, if we were to go through and roll them [the hotels] 200 times. I mean we could figure out what it would be compared to the houses.

Another pair thought they would need to toss 32 hotels (the number of houses available to the students in the sampling and probability task) five times, replicating their method of sampling with the houses during the sampling and probability task.

Since time constraints did not allow for replicating the sampling distribution and there were not 32 hotels available, the pairs chose a variety of sampling methods. One pair tossed 10 hotels two times, another pair tossed 10 hotels three times, and a third pair tossed 10 hotels 10 times, recording their results for the number of hotels landing upright. Three other pairs also tossed 10 hotels 10 times, averaging their results to obtain a proportion for the hotels landing upright. The seventh pair tossed one hotel 10 times, choosing that method to reduce the variability caused by the hotels bumping into one another as they were tossed. All of the pairs' tosses resulted in proportions that were at or close to the peak of the sampling distribution; however, their remarks demonstrated a variety of conclusions.

When students were drawing their conclusions about whether the probability that a hotel would land upright was the same as that for a house, three of the pairs (two who tossed 10 hotels 10 times and the pair tossing one hotel 10 times) thought the probability would be the same; however, one of those pairs was comparing their result of 0.17 to 0.19, their result from tossing the houses in the sampling and probability task. The remaining four pairs stated that they thought the probability would be less even though their tosses resulted in percentages that were at or close to the peak of the sampling distribution. They were not taking the natural variability that could occur into consideration and, therefore, were not appropriately using their data as evidence in drawing their conclusions. A summary of the number of tosses of hotels by each pair and their concluding remarks about whether the probabilities for hotels and houses landing upright were the same are displayed in Table 1.

Table 1 — Students' Concluding Remarks in Sampling Distribution Task.

Pairs and Number of Tosses	Students' Concluding Remarks
April and Brian 10 (averaged)	April: Like we had more two's and it looks like this one has more two's. So I feel like it would have the same probability as the house. Brian: I'm still doubtful. What we found was about 25%. So about a fourth of the time it'll land upright, if not a little bit more than that. And for this [the sampling distribution] we have like 20%, less than 20%, so that's just me doing math in my head and I just don't think it's likely. Plus we only did 10 trials.
Caitlin and David 2	David: I think you'd have to try probably more times, many more times, but, as it looks right now it's about the same. Caitlin: Maybe a little bit less.
Emily and Fritz 1 hotel 10 times	Fritz: Yeah, it was pretty similar. Between 1 and 2 [houses landing upright out of 10] so I think that [their results] still validates that that's relatively the same.
Gabrielle and Jared 3	Jared: I'm figuring it's not going to be that far away. I think it's going to be roughly the same. It's maybe just a little bit less because it's weighted differently.
Laura and Mark 10 (averaged)	Laura: So we got 17%. I don't remember how many we did for the last one but that's like fairly close. And I think we did more or we like threw more houses last time. So I think that's pretty... I'd say they are about the same. Mark: Yeah, I'd say about the same.
Rachel and Steve 10	Steve: We didn't do nearly enough. I mean you did this 200 times, we did this 10 times so like you can't really say like, oh look what we did really quick and that refutes that. Rachel: Then I'd say it's different, but not by a lot.
Nathan and Pete 10 (averaged)	Pete: You gotta take more samples. …but with one trial, I think regardless of the outcome, you can't really compare that to what you got from this population [referring to sampling distribution]. It may fit into what you have seen. Like right here, this value right here, like 1, 2, [referring to peak in sampling distribution] ours was close so we could say yeah, it does compare similarly but I'm not going to bet my life on it.

The majority of the pairs demonstrated difficulty in making an accurate informal inference even though they had identified likely and rare outcomes with the sampling distributions of mean areas of rectangles in part two of the task. Additionally, the students had recently studied the normal distribution in their statistics classes which included the 68-95-99.7 rule of percentages of data within one, two, and three standard deviations of the mean. For the majority of them, this did not translate into the variability associated

with this sampling distribution or the understanding that they could draw a conclusion based on a single sample of data.

Returning to the Makar and Rubin (2009) framework for thinking about informal statistical inference, all students were able to make an inference based on the data; however, the majority of them did not believe they had enough data to draw a conclusion about the probabilities. They demonstrated their probabilistic reasoning with phrases such as "relatively the same", "maybe a little bit less", "it's different, but not by a lot", or "I'm not going to bet my life on it." All students also used their data as evidence for making their inference with the majority of them comparing proportions rather than considering their results in relation to the sampling distribution presented to them.

At least one student from four of the seven pairs stated that the probability of a hotel landing upright was slightly different than that of a house. Additionally, four of these pairs expressed skepticism in the accuracy of their results due to their small number of tosses. The majority of these students were not yet ready to draw a conclusion from a single sample of data using the variability of the sampling distribution.

However, their statements, many expressing a degree of certainty or uncertainty, provided evidence that they were at a point in their informal inferential reasoning when they might be able to consider this next level of reasoning. This could be seen when the following question was asked of the pair who tossed 10 hotels three times:

> Interviewer: So is it [the results] enough less to say, do you think, that the probability is different?
>
> Jared: How far away would it have to be? Like, I mean, I don't know, I think it would be a couple percentage points. You know just a little lower.

This student's question expressed the foundation of formal statistical inference. We take the posing of this question as an indication that this student's understanding of what it means for these sample proportions to be relatively the same or slightly different is still developing. The impact of this statistical reasoning still in development could be seen in students' responses during the formal statistical inference task.

The Formal Statistical Inference Task

While constructing a confidence interval, students wanted to take several samples for their point estimate. Following this, students demonstrated their procedural knowledge of formal statistical inference rather than a conceptual knowledge. They were able to construct a confidence interval and conduct a hypothesis test; however, they were uncertain of the meaning of the confidence level and the *p*-value.

In forming their own confidence interval for the proportion of red beads in the container of multi-colored beads, two pairs took 10 samples to find the sample proportion, *p*-hat, and one pair took three samples. These students were averaging to obtain their value for the sample proportion to use as their estimate, not believing that one sample proportion would provide accuracy. This could be seen when Caitlin and David began to

work on their confidence interval by taking more than one sample of 32 beads with the sampling device.

Interviewer: So how many samples will you need to construct a confidence interval?

David: You should have a large number, [pause] but actually you wouldn't need many once you get your original proportion.

Caitlin: You could use just one.

David: You could use one.

Caitlin: That's all you really need.

When pressed to think about the components of a confidence interval, Caitlin and David realized they needed just one sample proportion. April and Brian, however, could not be convinced.

Brian: Eight, maybe 8 times doing this [sampling] will be close to 2000 [total number of beads in the container], no it wouldn't. Eighty times doing this would be close to 2000 beads. But I don't want to do this 80 times. Um.

Interviewer: Yeah, how many times do you need to do it [take samples for the sample proportion]?

Brian: Well the more we do it [take sample proportions] the better it is [the sample proportion for the estimate in the confidence interval].

Despite being questioned by the interviewer, Brian and April took 10 sample proportions of red beads and averaged them to obtain the sample proportion for their interval. Brian alluded to the long-run characteristic of probability when he said that their sample proportion for the estimate in their confidence interval would be better with more trials. This was an indication that Brian was trying to minimize the variability in their estimate for the sample proportion. This pair was either not aware of the power of the sampling distribution used to formulate the confidence interval or unwilling to rely on that power.

Throughout the formal statistical inference task, the pairs of students demonstrated their procedural knowledge in constructing confidence intervals and conducting hypothesis tests. However, when asked about the meaning of the confidence level, the pairs of students did not demonstrate an understanding that this value was based on the sampling distribution and its normality. For example, three of the pairs interpreted the 90% in their confidence interval as the chance the true population proportion of red beads would be in the interval; one pair interpreted it as the percentage of intervals constructed that would result in the exact same interval; two of the pairs stated that they did not know how to interpret the 90%; and the remaining pair was the only one to allude to 90% of multiple trials, but could not provide a complete explanation. When asked to define the *p*-value in their hypothesis test, six of the pairs described the procedure

of comparing it to the significance level for drawing a conclusion. The remaining pair stated that the *p*-value was a probability, but neither student could state to what this probability referred.

During the sampling distribution task, students demonstrated an understanding of the characteristics of the sampling distribution which included the fact that it is approximately normal and that the larger the sample size is the less the variability in the sampling distribution. They could identify likely and unlikely sample means when viewing a series of sampling distributions of random rectangles. However, when asked to draw a conclusion about whether the proportion of Monopoly hotels landing upright when tossed was the same as that for the houses based on a sampling distribution, students were not confident that they could draw a conclusion based on a single sample and wanted to take several samples before drawing a conclusion. This was an indication that the students were not yet understanding the role and the power of the sampling distribution in statistical inference; and, therefore, their conceptual understanding of statistical inference was still developing. During the formal statistical inference task, students also exhibited the propensity to take several samples rather than a single sample to form a confidence interval and then demonstrated procedural rather than conceptual knowledge while completing the formal statistical inference task.

Discussion

In completing the last part of the sampling distribution task, understanding of the characteristics of the sampling distribution did not translate into determining likelihood based on a single sample proportion. Characteristics of their reasoning appeared to be based on reasonings they demonstrated during the comparing distributions task and the sampling and probability task. Students wanted to compare the sampling distribution for the proportion of Monopoly hotels landing upright when tossed to that of the proportion of houses to draw a conclusion as they had done in the comparing distributions task. While they did not ultimately generate the sampling distribution for the proportion of hotels landing upright, the majority of them did take several samples exhibiting an adherence to the long-run characteristic of probability. Their lack of a fully developed understanding of the power of the sampling distribution in statistical inference was then demonstrated in their inability to give a meaningful interpretation of the confidence level in a confidence interval and *p*-value in the formal statistical inference task.

Informal Inferential Reasoning with the Sampling Distribution

When drawing informal inferences with the sampling distribution in the last part of the sampling distribution task, the majority of the students exhibited uneasiness in relying on a single sample of data. Overall the students had a general knowledge of the sampling distribution. They knew it took on the shape of a normal distribution and they made references to the decrease in variability as the sample size increased. They

also identified sample statistics that would be considered likely or rare based on the probabilities associated with the normality of sampling distributions. However, despite this general knowledge, rather than relying on a single sample proportion of Monopoly hotels landing upright, students wanted to construct a sampling distribution for the proportion of Monopoly hotels landing upright. This sampling distribution could then be compared to the sampling distribution for the proportion of houses landing upright to draw a conclusion about whether those proportions were the same.

In light of the research on students' understandings of the sampling distribution, Saldanha and Thompson (2002) found the majority of their students compared a single sample statistic to the population parameter rather than to the sampling distribution of all such statistics when asked to determine if it was unusual. When the introductory students in this study were asked if they could determine if the probability that a hotel would land upright was the same as that for a house, they were also hesitant to compare their sample proportion to the sampling distribution of proportions presented to them. However, these students were not able to compare to a population parameter as were the students in Saldanha and Thompson's study since the true p was not known and there was no clear theoretical distribution for the population proportion of houses landing upright when tossed.

Saldanha and Thompson (2002) also found that students were able to reason about the variability among samples; however, this reasoning did not necessarily translate to the variation that existed in sample statistics. The students in our study wanted to take many samples, exhibiting similar reasoning, to diminish the effects of the variability that could occur between samples. Like the students in Saldanha and Thompson's study, this understanding of variability did not translate to the variability associated with the sampling distribution of sample statistics. This prevented them from having confidence that they could draw a conclusion based on just one or even a small number of samples. Taking many samples with the hotels may have felt like a much more concrete method for inferring about the probability that a hotel would land upright. Possibly relying on a sampling distribution that they did not create and could not be sure of the circumstances under which it was created, was viewed as an uncertainty to avoid.

Pratt and colleagues (Pratt *et al.*, 2008) examined local and global thinking as students reasoned informally using software that gave them the ability to add to an existing sample or generate a new sample. In either situation, students tended to focus on the changes in subsequent displays of the data. At times when they did express a global understanding by referring to the stability found when considering all of the samples, they were still frustrated by the fluctuations they saw in the individual samples. This suggested that an important aspect of informal inference is in finding the invariance that is present even among all of the local changes. The majority of the students in our study viewed their samples in a local sense inasmuch as they were concerned whether each sample was accurate enough to draw a conclusion. For several of the students, a result of one out of 10 hotels landing upright, for example, triggered a response that this result differed from the two out of 10 peak of the sampling distribution. They were not viewing

their samples in a global sense when comparing them to the sampling distribution; and, therefore, could not rely on the variability associated with its normality.

Informal Inferential Reasoning with the Sampling Distribution and Prior Informal Inferential Reasonings

The characteristics of the statistical reasoning students exhibited in the sampling distribution task appeared in their reasoning in the first two tasks. Some (but not all) of these characteristics are carried through to the final task on formal statistical inference. We take this as an indication that the students' reasoning had not yet fully developed so as to tackle the intricacies of drawing conclusions from the sampling distribution.

Based on their responses, the majority of the students wanted to return to comparing distributions by generating a sampling distribution for the proportion of hotels landing upright rather than use the given sampling distribution for the proportion of houses landing upright (shown in Figure 4). We interpreted this as the comparing of measures of center in distributions, similar to what they had done in the comparing distributions task, provided these students with a higher degree of certainty than relying on the normality of the sampling distribution. Therefore, comparing two sampling distributions was a viable option for determining if the hotels behaved similar to the houses. When comparing distributions of class test scores in the sampling distributions task, students primarily focused on the mean or median in drawing a conclusion as to which class scored better on a test (Jacob, 2013). For the majority of these students, as they compared symmetrical distributions with the same mean and median, the equitable measures of center were more important than the differences in variability in drawing a conclusion. Many of them referred to the differences in variability between the distributions; however, did not see that as helpful in drawing a conclusion. This presented a barrier for them in using the sampling distribution to draw a conclusion.

The long-run characteristic of probability appeared to have an impact on these students as well. During the sampling and probability task, when students took their own samples to approximate the proportion of houses landing upright when tossed and to approximate the proportion of green beads in the container, they demonstrated a clear understanding of the long-run characteristic of probability (Jacob, 2013). Students had difficulty in transferring this concept to the sampling distribution which was generated from a large number of samples. They held on to their accurate concept of the long-run characteristic of probability in one sense, wanting to take many more samples to compare to the sampling distribution. While this was an attempt to reduce variability, which is a correct intuition, using the known variability that exists in the sampling distribution is a more efficient method. This same type of reasoning appeared during the formal statistical inference task when many of these students wanted to take several samples for the proportion estimate to construct a confidence interval for the proportion of red beads in the container.

Informal Inferential Reasoning with the Sampling Distribution and Formal Inferential Reasoning

Whether students' difficulties in relying on the sampling distribution stemmed from not considering the sampling distribution a powerful tool generated from a large number of samples or whether they found more certainty in comparing measures of center, this difficulty prevented them from making the necessary connections for formal statistical inference. This difficulty appeared during the formal statistical inference task when three of the student pairs took several samples for the proportion estimate to construct a confidence interval. This was not incorrect statistical reasoning; they were trying to obtain a more accurate estimate by using several samples. As the formal statistical inference task progressed, however, it was evident that these students were primarily depending upon their procedural knowledge. Without a deep understanding of the power of the sampling distribution, relying on it for formal statistical inference was not possible. Therefore, as students learned the procedures for formal statistical inference, the role of the sampling distribution was likely not paramount. Students appeared to have made little connection between the procedures for formal statistical inference and the underlying conception of the sampling distribution. This was evidenced as all seven pairs were successful in creating a confidence interval and conducting a hypothesis test based on their own samples. However, when asked about the meaning of the confidence level and the *p*-value, they did not demonstrate an understanding that these values relied on the sampling distribution and its normality. Without this connection, they could not give meaningful interpretations of the confidence level or the *p*-value; and instead relied on formulaic knowledge and correctly memorized formal statements of conclusions.

Implications for Practice and Conclusions

For the students in this study, it appeared that their statistical reasoning in working with the sampling distribution was still in development rather than entirely correct or incorrect. Developing activities that allow introductory statistics students to explore their notions regarding samples and comparing them to the sampling distribution may support their informal inferential reasoning. For example, if they were able to generate another sampling distribution for comparison as many of the pairs initially wanted, this may help them to understand that taking a larger number of samples will not necessarily provide more certainty in their conclusions. Discussions about efficiency as well as accuracy in data collection might help students develop their statistical reasoning. Introductory statistics students likely do not have any practical experience working within budget or time constraints for using data to draw conclusions; therefore, creating a sampling distribution and/or taking many samples to compare and then make a decision may have seemed like a reasonable method. In addition to giving students more experiences with efficiency in data collection and drawing conclusions, more time could be spent on inferring with the sampling distribution before introducing the procedures of formal statistical inference. This may help students to see the need for the significance level in

hypothesis testing as the "cut off" between reasonably likely and unlikely. This may also enhance their understanding of the significance of *p*-values in formal statistical inference. Students likely need multiple experiences to help them understand that one sample can be enough to draw an inference about a population.

References

American Statistical Association. (2010). *Guidelines for assessment and instruction in statistics education (GAISE) college report*. Retrieved from http://www.amstat.org /education/gaise

Bakker, A., & Gravemeijer, K. (2004). Learning to reason about distribution. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 147–168). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Ben-Zvi, D. (2004). Reasoning about variability in comparing distributions. *Statistics Education Research Journal*, *3*(2), 42–63. Retrieved from http://www.stat.auckland.ac.nz /~iase/serj/ SERJ3(2)_ BenZvi.pdf

Castro Sotos, A. E., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2009). How confident are students in their misconceptions about hypothesis tests. *Journal of Statistics Education*, *17*(2). Retrieved from http://www.amstat.org/publications/jse/v17n2/castrosotos.html

Chance, B., delMas, R., & Garfield, J. (2004). Reasoning about sampling distributions. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 295–323). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Finzer, W. (2001). *Fathom Dynamic Data Software*. (Version 2.1) [Computer Software]. Emeryville, CA: Key Curriculum Press.

Garfield, J., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice*. New York, NY: Springer.

Goldin, G. A. (2000). A scientific perspective on structured, task-based interviews in mathematics education research. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 517–545). Mahwah, NJ: Laurence Erlbaum Associates.

Jacob, B. L. (2013). *The development of introductory statistics students' informal inferential reasoning and its relationship to formal inferential reasoning*. Unpublished doctoral dissertation, University of Syracuse, New York, United States of America. Retrieved from http://surface.syr.edu/tl_etd/245

Kelly, B. A., & Watson, J. M. (2002). Variation in a chance sampling setting: The lollies task. In B. Barton, K. C. Irwin, M. Pfannkuch, & M. O. J. Thomas (Eds.), *Mathematics education in the South Pacific* (Proceedings of the 25th annual conference of the Mathematics Education Research Group of Australasia, Vol. 2, pp. 366–373). Sydney, NSW: MERGA.

Konold, C. (1989). Informal conceptions of probability. *Cognition and Instruction*, *6*(1), 59–98.

Konold, C., Madden, S., Pollatsek, A., Pfannkuch, M., Wild, C., Ziedins, I., Finzer, W., Horton, N. J., & Kazak, S. (2011). Conceptual challenges in coordinating theoretical and data-centered estimates of probability. *Mathematical Thinking and Learning*, *13*(1), 68–86. Doi: 10.1080/10986065.2011.538299

Makar, K., & Confrey, J. (2005). "Variation-talk": Articulating meaning in statistics. *Statistics Education Research Journal*, *4*(1), 27–54. Retrieved from http://www.stat.auckland.ac.nz/~iase/serj/ SERJ4(1)_Makar_Confrey.pdf

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal*, *8*(1), 82–105. Retrieved from http://www.stat.auckland.ac.nz/~iase/ serj/SERJ8(1)_Makar_ Rubin.pdf

Pfannkuch, M. (2006). Comparing box plot distributions: A teacher's reasoning. *Statistics Education Research Journal*, 5(2), 27–45. Retrieved from http://www.stat.auckland.ac.nz/~iase/serj/SERJ7(2)_Pfannkuch.pdf

Pratt, D., Johnston-Wilder, P., Ainley, J., & Mason, J. (2008). Local and global thinking in statistical inference. *Statistics Education Research Journal*, *7*(2), 107–129. Retrieved from http://www.stat.auckland.ac.nz/~iase/serj/ SERJ7(2)_ Pratt.pdf

Reading, C., & Shaughnessy, M. (2000). Student perceptions of variation in a sampling situation. In T. Nakahara & M. Koyama (Eds.), *Proceedings of the 24th Conference of the International Group for the Psychology of Mathematics Education* (pp. 89–96). Hiroshima, Japan: Hiroshima University.

Saldanha, L., & Thompson, P. (2002). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics*, *51*(3), 257–270.

Shaughnessy, J. M., Canada, D., & Ciancetta, M. (2003). Middle school students' thinking about variability in repeated trials: A cross-task comparison. In N. A. Pateman, B. J. Dougherty, & J. T. Zilliox (Eds.), *Proceedings of the 27th Conference of the International Group for the Psychology of Mathematics Education* (pp. 159–165). Honolulu, HI: Center for Research and Development Group, University of Hawaii.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Watson, J. M. (2002). Inferential reasoning and the influence of cognitive conflict. *Educational Studies in Mathematics*, *51*(3), 225–256.

Watson, J. M., & Moritz, J. B. (1999). The beginning of statistical inference: Comparing two data sets. *Educational Studies in Mathematics*, *37*(2), 145–168.

Zieffler, A., Garfield, J., Delmas, R., & Reading, C. (2008). A framework to support research on informal inferential reasoning. *Statistics Education Research Journal*, *7*(2), 40–58. Retrieved from http://www.stat.auckland.ac.nz/~iase/serj/SERJ7(2)_Zieffler.pdf

Resumo. O raciocínio estatístico em torno da distribuição amostral é necessário para a inferência estatística formal. O nosso estudo, com alunos de uma disciplina introdutória de estatística com idades entre 16-18, sugere que o conhecimento das características de uma distribuição amostral e as experiências com a geração de distribuições amostrais não fornecem uma base suficiente para este raciocínio. Após instrução sobre a distribuição amostral e as suas características, os alunos neste estudo tiveram oportunidade de obter uma amostra antes de fazer uma inferência informal baseada numa distribuição amostral. A maioria dos alunos utilizou várias amostras e/ou considerou gerar uma segunda distribuição amostral para comparação. Estas noções estatísticas não são incorretas; pelo contrário, indicam que o seu raciocínio estatístico sobre a distribuição amostral ainda está em desenvolvimento. Os resultados são discutidos em relação ao raciocínio inferencial informal prévio dos alunos e o seu posterior raciocínio inferencial formal.

Palavras chave: Educação estatística; distribuição amostral; raciocínio inferencial.

Abstract. Statistical reasoning surrounding the sampling distribution is necessary for formal statistical inference. Our study of introductory statistics students aged 16-18 suggests that knowledge of the characteristics of a sampling distribution and experiences with generating sampling distributions do not provide a sufficient basis for this reasoning. Following instruction on the sampling distribution and its characteristics, the students in this study were given the opportunity to draw a sample before making an informal inference based on a sampling distribution. The majority of the students took several samples

and/or considered generating a second sampling distribution for comparison. These are not incorrect statistical notions; but instead indicate that their statistical reasoning with the sampling distribution was still developing. The results are discussed in relation to students' prior informal inferential reasoning and their subsequent formal inferential reasoning.

Keywords: Statistics education; sampling distribution; inferential reasoning.

■■■

BRIDGETTE JACOB
Onondaga Community College, Syracuse University, New York
jacobb@sunyocc.edu

HELEN M. DOERR
Syracuse University, New York
hmdoerr@syr.edu

(Recebido em março de 2014, aceite para publicação em junho de 2014)