Exploring students' models of sampling and inference with nominal variables

Explorando os modelos de estudantes sobre amostragem e inferência com variáveis nominais

Jonas Bergman Ärlebäck
Department of Mathematics, Linköping University
Sweden
jonas.bergman.arleback@liu.se

Peter Frejd
Department of Mathematics, Linköping University
Sweden
peter.frejd@liu.se

Helen M. Doerr
Department of Mathematics, Syracuse University
USA
hmdoerr@syr.edu

Abstract. Making inferences about unknown populations is central in statistical reasoning. However, little attention has been paid to empirical investigations of how and why students develop sampling models when investigating a categorical variable whose values are nominal. This paper reports on an intervention that draws on the models and modeling perspective, where 25 pre-service teachers were asked to develop a sampling model that could be used to make inferences about the number of different colored beads and the distribution of different colored beads in different sized populations. Using a thematic analysis, three main results about the characteristics of the students' models of sampling and inference with nominal variables were identified: to catch all the low frequency colors in the population; not to overestimate the low frequency colors in the population; and, to formalize the relationships used in making inferences. These results highlight several issues about students' understanding of the relationship between sample representativeness and sample variability and its consequences for making inferences.

Quadrante: Revista de Investigação em Educação Matemática 30(1) 158-177
https://doi.org/10.48489/quadrante.23655
Received: February 2021 / Accepted: May 2021

Keywords: categorical variables; eliciting students' ideas; models and modeling perspective; nominal values; pre-service teachers; sampling.

Resumo. Fazer inferências sobre populações desconhecidas é fundamental no raciocínio estatístico. No entanto, pouca atenção tem sido dada às investigações empíricas sobre como e com que objetivo os alunos desenvolvem modelos de amostragem ao investigarem uma variável categórica cujos valores são nominais. Tendo por base a perspetiva de modelos e modelação, este artigo relata uma intervenção, durante a qual 25 professores em formação inicial foram solicitados a desenvolver um modelo de amostragem que pudesse ser usado para fazer inferências sobre o número de contas de cores diferentes e sobre a distribuição de contas de cores diferentes em populações de tamanhos diferentes. Por meio de uma análise temática, foram identificados três resultados principais sobre as características dos modelos dos estudantes acerca de amostragem e inferência com variáveis nominais: capturar todas as cores de baixa frequência na população; não sobrevalorizar as cores de baixa frequência na população; e formalizar as relações encontradas para fazer inferências. Os resultados põem em evidência várias questões sobre a compreensão dos alunos acerca da relação entre a representatividade da amostra e a variabilidade da amostra e suas consequências na produção de inferências.

Palavras-chave: variáveis categóricas; suscitar as ideias dos alunos; perspetiva de modelos e modelação; categorias nominais; formação inicial de professores; amostragem.

Introduction

The abilities to interpret and make sense of data in various forms in people's private and professional lives are now more important than ever before (Manyika et al., 2011). However, data itself does not tell us anything, but requires being organized, examined and reasoned with using various models in order to provide us with information and knowledge. Hence, modeling and statistical reasoning have become increasingly important for students at the secondary and post-secondary levels. Students need to understand and become proficient with a wide range of statistical concepts and learn the forms of reasoning that are needed to interpret and make sense of data (Franklin et al., 2007; OECD, 2013). Ben-Zvi, Bakker, and Makar (2015) stress that the core of statistical reasoning is to make inferences about unknown populations using samples of data that are representative of the population. Thus, adequate models of sampling are crucial in order to make sound inferences about a population.

The research that has been conducted to date with respect to students' thinking and approaches to sampling has been focused primarily on using sampling models to make inferences about population statistics such as mean, median, mode and proportions (Batanero et al., 2020; Lipson, 2003; Saldanha & Thompson, 2002). The population statistics in these studies are derived from different types of variables describing various attributes of the population. There are many ways to define different types of variables satisfying a plethora of conditions and suitable for various purposes (Dodge, 2008; Everitt & Skrondal, 2010). The type of a variable is often strongly connected to the type of data it represents and common types are interval, ratio, ordinal and nominal variables.

Students' thinking about sampling models for making inferences about the number of categorical values and the distribution of a nominal variable in a population has been given little attention in the research. According to Budgett and Puloka (2019), reasoning with categorical data is challenging for both students and teachers, with their understandings and use of proportional reasoning being a major confounding factor (Batanero et al., 1996; Böcherer-Linder et al., 2018; Watson & Callingham, 2014). In this paper, our goal is to explore the sampling models students develop when engaged in a sense-making modeling task about making inferences about the attributes of a nominal variable for different size populations.

Previous research

Discussing the process of making inferences about unknown populations using samples of data in general, Ben-Zvi et al. (2015) stress that research points to the importance of the samples used being representative of the population in question. The samples should exhibit a balance between sample representativeness and sample variability (Saldanha & Thompson, 2002; Shaughnessy, 2007). To overly rely on one or the other might result in the belief that the samples tell you either absolutely everything or absolutely nothing about the population (Ben-Zvi et al., 2015).

Another approach discussed in the literature on sample representativeness and sample variability is the idea of bootstrapping. Bootstrapping is a collective name of a family of repeated random sampling approaches usually involving replacement. Multiple researchers (e.g., Engel, 2010; Hesterberg, 2006) have pointed out the benefits of bootstrapping methods to understanding key concepts in statistics such as the logic of inference. However, McLean (2015) has noted that there is limited research on students' reasoning about sampling and making inferences using bootstrapping.

Although the research focusing on sampling methods for making inferences about the number of categorical values and the distribution of a nominal variable in a population is sparse, a few related research studies are found in the literature. These studies almost exclusively focus on bivariate categorical data that can be represented in two-way tables. Budgett and Puloka (2019) investigated the questions posed and the ways undergraduate students reasoned based on different representations of bivariate categorical data. These researchers found that working with situations involving categorical data and making comparisons were ambiguous for the students. Whether the data were presented graphically or in two-way tables, the students struggled with doing within-category and between-category comparisons, often leading to conflicting conclusions and inferences.

Budgett and Puloka (2019) noted that different ways to parse the categorical data and understanding the associations between the variables can be seen as different conditional situations, which influenced how students connected and associated the categorical data.

When investigating 11-13 year old students' reasoning about associations of categorical variables, Casey et al. (2018b) found similar results. Neither data tables nor various graphs representing different conditions provided support for the students in thinking productively about the categorical data and assessing potential associations between the categories. These results and challenges persist throughout the educational system and apply to pre- and in-service secondary mathematics teachers (Casey et al., 2018a). These researchers found that instruction focused on graphical representations, proportionality and percentages was successful in developing teachers' abilities to more adequately reason with bivariate categorical data. The research we present in this paper complements and extends this earlier research by focusing on students' ideas and their developing sampling models as they seek to make inferences about a categorical variable with nominal values.

Theoretical framework

In our research study, we drew on a models and modeling perspective (Lesh & Doerr, 2003). Following Doerr and English (2003), we understand a model to be a "system of elements, operations, relationships, and rules that can be used to describe, explain, or predict the behavior of some other familiar system" (p. 112). Models are not stand alone solutions to single and isolated problems, but rather generalizable systems that can be applied to similar problems and situations (Lesh & Doerr, 2003). Our use of this framework for our task design and data analysis is in line with Doerr et al. (2017), who argue that a models and modeling perspective incorporates key tenets from modeling perspectives in science education (Hestenes, 2010), statistics modeling (Lehrer & Schauble, 2010) and mathematical modeling (Barbosa, 2006). The models and modeling perspective stresses that models are iteratively developed generalizable sense-making systems that create shared representations and meaning of the phenomena investigated. In this perspective, the first step in supporting students in developing a model to make sense of a situation or problem is often using a model eliciting activity (MEA) (Lesh & Doerr, 2003). The purpose of an MEA is to elicit the students' initial ideas about the problem situation to make these explicit and to facilitate discussion, evaluation and further development. The activity used in this study was specifically designed to investigate the models elicited when students engaged in a sense-making activity involving categorical variables taking nominal values given a physical problem situation and hands-on materials.

In our study, we examined the models that students developed to make inferences about the number of categorical values and the distribution of a nominal variable in a population. Using the models and modeling perspective, we conceptualized the models that students developed to consist of elements in terms of the actual samples collected, the quantities derived from the samples, and the conceived relationships among the collected samples, the derived quantities and the population at large. The collected samples are taken based on rules that determine a sampling procedure using attributes of the samples such as sample size and the number of samples. These attributes can be manipulated mathematically using specified operations resulting in derived quantities, relationships and representations which are used to make inferences about given populations. Thus, students' models of sampling and inference consist of the rules that determine the sampling procedure, the mathematical operations used on the samples and how these are used in the inferences made by the students based on their conceived relationships between the samples and the populations.

Aim and research questions

The aim of this paper is to explore the sampling models students develop and the inferences made when investigating a categorical variable whose values are nominal in three different sized populations. To this end, we address the following research questions:

1. What relationships did the students focus on to determine if their sampling models were reasonable for making an inference about a particular population?
2. How and why did the students develop rules about sample size, number of samples and replacement?

Setting and analysis

To address our research questions, we designed a task aimed at eliciting the students' sampling models. We analyzed the work of eight groups of preservice secondary teachers (from here on referred to as students) on the task using an exploratory thematic analysis. We will elaborate on the task and on the analysis below, but first we describe the research setting.

Setting and data collection

The 25 participating students were enrolled in a secondary mathematics teacher training program. Prior to participating in the study, these students had all taken at least one semester of mathematics courses on topics such as calculus, linear algebra, statistics and mathematics education. The statistics course was an eight-week course comprised of 12 formal lectures presenting the content and theory and 12 lessons where the students worked on assigned problems connected to the content. The course was mathematically and theoretically oriented, and the content covered stochastic variables, probability distributions, normal and binomial distributions and formal inference. The role and function of samples were strongly tied to the theory about stochastic variables.

In our study, the students were divided into eight groups of two to four students and placed in separate rooms where they were recorded with either a camera or an audio-recorder. We asked students to think out loud during their work on the task. In the rooms, the students and the available materials were positioned so that the group was captured by the fixed camera or within the uptake radius of the audio-recorder. The first two authors circulated among the rooms making sure that the recording devices functioned properly and that the students understood the task. The data collected to document the students' work consisted of audio recordings (for two groups) and video recordings (for six groups) and all individual students' written work. The students all had individual worksheets on which they documented to varying degrees their work. All groups designated one member to take more careful notes of their collective work. The groups worked between 90 and 120 minutes on the task.

The task and its design

To investigate the sampling models that the students developed, we created physical populations with a nominal attribute (color) that was readily visible to the students. We created three different sized populations of colored beads (N = 20, 400, 10 000). For each size population, we created two populations with visibly different distributions of colored beads: one nearly uniform and one clearly skewed. This resulted in a total of six populations for the students to investigate and develop a sampling model for making inferences (see Figure 1). Each different color bead represented a particular category. Hence, the distribution of colors in each container corresponded to the distribution of a nominal variable.

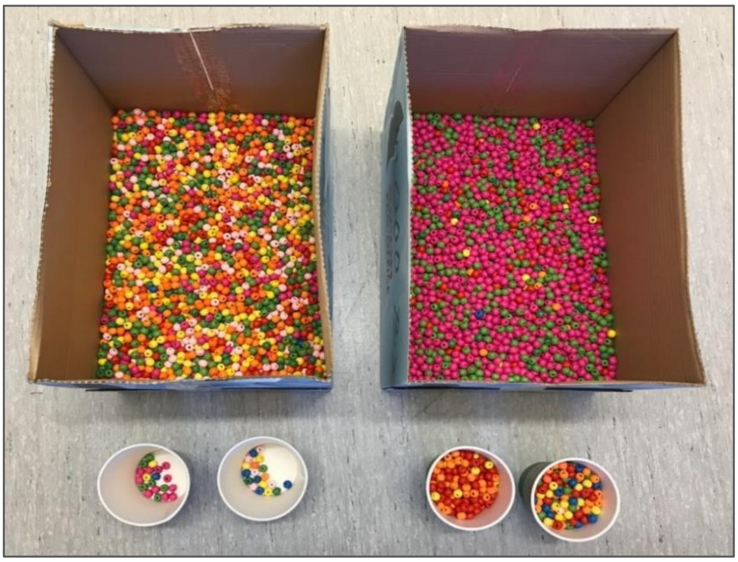

Figure 1. The six containers of beads (two each of N = 20, 400, 10 000)

Drawing on the models and modeling perspective, the students were asked to develop a sampling model that could be used to make predictions about the number of different colored beads and the distribution of different colored beads in any size population (see Figure 2). Note that the first question (Q1) asks the students to make an inference about a single variable, which is a single-valued statistic, namely the number of different colors in a population. On the other hand, the second question (Q2) asks for an inference with respect to the distribution of colors in a given population, which is a multi-valued statistic.

> *Develop a sampling model to investigate the following two questions:*
>
> *Q1. How many different colors of beads are there in a given container containing N beads?*
>
> *Q2. What is the distribution of colored beads in a given container containing N beads?*
>
> *The sampling model you develop should be applicable to any container containing an arbitrary (but known) number of beads. Test you sampling model on the six provided containers containing N different colored beads (N=20, 400, 10 000; two containers of each size) and reflect on your sampling model. Document your work on the provided worksheet.*

Figure 2. Students' task for making inferences about a nominal variable

The design of the task was based on the following five guiding principles (partly inspired by, but not identically to, the design principles of MEAs): First, we wanted the task to engage the students to the exploration and sense-making of an actual sampling situation. Hence, we decided to provide the students with physical material to use for their sampling, rather than use pre-defined samples or a computer simulation to generate samples (see Figure (3a)). Second, we wanted the sampling models developed by the students to be general and applicable to a general population. Hence, as shown in Figure 1, we (a) provided each group of students with six containers of three different size populations (*N* = 20, 400, 10 000) with the population size explicitly written on the containers, and (b) prepared the distributions of the two containers with equal population sizes so that one was roughly uniform and one was skewed. Third, to explicitly focus the students' attention on the possibility that the distributions of the populations they investigated could be skewed, we provided a seventh container with 20 beads which the students could use in any way they wanted. The distribution of colors in the seventh container was skewed with nine red beads, five green beads, five orange beads, and one yellow bead, as shown in Figure 3b. Fourth, to facilitate the actual sampling of the beads, we provided a sampling device consisting of a plastic pipe of suitable diameter cut in half length-wised with closed ends (see Figure (3b)). This sampling device could hold a single maximum sample between 20 and 24 beads. Fifth, we included a closed envelope with data on the actual bead distributions of the containers. We instructed the students to open this envelope at the conclusion of their activity on the task. We intended the data on the actual distributions to be used as a reference in discussing and writing about the strengths and weaknesses of their sampling models.

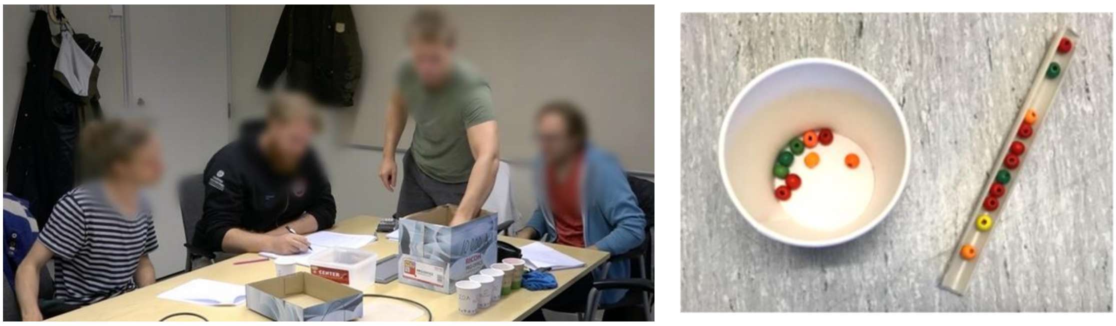

Figure 3. (a) Students developing and enacting their sampling model; (b) The seventh container of beads and the sampling device

Analysis

We analyzed the data in three stages using thematic analysis (Braun & Clarke, 2006). We began with a theoretically based inductive analysis followed by an open coding process to create summaries of the audio and video data. The students' written work documenting the representations of their developing sampling models was also analyzed and used for clarifying ambiguities in the audio and video material. The themes we identified, reflecting the main patterns in our data, provided the answers to our research questions. The analysis was carried out collaboratively by the three authors as outlined below.

In the first stage of the analysis, we created detailed written summaries of the groups' work on the task based primarily on the audio and video data. The students' written work was used as a secondary source to clarify ambiguities in the audio and video data. Rather than including the students' exchanges word for word, the summaries mixed quotations of students' utterances with descriptions of actions and interactions within the groups. This provided us with a coherent and continuous narrative of each group's work. In creating the summaries, we drew on the models and modeling perspective (Doerr & English, 2003; Lesh & Doerr, 2003) to make sure all instances of discussion and work involving samples, sampling, mathematical operations on collected samples and making inferences were captured and described in the summaries. This stage of the analysis resulted in five to eight page summaries of each of the eight groups of students. Given that the students spoke Swedish, the first two authors made the summaries and translated these into English, which then were discussed for clarity and coherence among all three authors.

In the second stage of the analysis, we used a priori codes in terms of elements, relationships, rules, and operations from the models and modeling perspective to examine the sampling models that the students' developed. This analysis parsed the summaries into segments focusing on different aspects of the students' sampling models and how those aspects developed.

In the third stage of the analysis, we applied inductive open-coding to the segments identified in stage two focusing on characterizing the relationships and rules in the students' emerging sampling models. We focused on identifying the main characteristics of the students' sampling models and describing students' reasoning as to why and how to make inferences for different sized populations about distributions of a categorical variable whose values are nominal. This analysis resulted in the themes that are reported in the results section below. In both stage two and three of the analysis, the coding was made inductively, iteratively, and collaboratively by the three authors and discrepancies were discussed and resolved as an integral part of the coding process.

Results

We identified three major themes about the characteristics of the students' models of sampling and inference about nominal variables: to catch all the low frequency colors in the population; not to overestimate the low frequency colors in the population; and to formalize the relationships facilitating inferences. These three themes will be used to describe the relationships that the students focused on to determine if their sampling models were reasonable to make an inference and to describe how and why the students developed rules about sample size, number of samples and replacement.

Catch the low frequency colors

The seventh container included 20 beads with only one yellow bead that was clearly visible to the students. When they did not catch the yellow bead in testing their sampling models, many groups of students expressed concern, especially with respect to the first question (Q1) in the task about identifying the number of different colors in the population. Several groups considered it important for their collection of samples to contain the low frequency colors in order to justify that their sampling models were reasonable for making an inference about the population. We illustrate the students' approaches to capturing the outliers (that is, colors of beads that occurred with low frequency in the population) with the work of three different groups.

The first example is from the work of the two students (S1, S2) in Group A and highlights how the students used multiple repeated samples with replacement to address their concern about catching the low frequency colors. In addition, this example illustrates how the students' sampling model developed by introducing a new rule deciding on the number of samples to collect. Group A's initial sampling model contained the rule of taking two samples of five beads with replacement (replacing the first sample back in the population before taking the second sample). When the students tested their model on the seventh container and compared their result with the actual content of the container, they realized that they did not catch all the colors. They then decided to adapt a new rule to determine the number of colors in a population by taking repeated samples (with replacement) of five beads until no previously yet unseen color appeared in their sequence of samples. The students' implementation of the new rule on another and nearly uniform 20-bead container is displayed in Table 1.

Table 1. Four samples with replacement collected by Group A to evaluate their developing sampling model.

Sample 1	Sample 2	Sample 3	Sample 4
2 blue	1 blue	1 blue	1 blue
1 yellow	2 yellow	1 green	2 green
2 green	1 green	**1 orange** (new color)	1 yellow
	1 pink (new color)	2 pink	1 orange

Table 1 shows that the students got a new colored bead in Sample 2 (pink) and another new color in Sample 3 (orange), but not in Sample 4 and therefore they stop taking more samples. Altogether they collected beads of five different colors, which they found a satisfactory result, since they could see by direct inspection that five was the correct number of colors in the container. However, testing their sampling model with the skewed distributed 20-bead container and skewed distributed 400-bead container, and by making a direct comparison to the actual distributions in the containers, did not given them satisfactory results. This prompted a discussion about the "naivete" (S1) of their approach. Student S2 commented that "you get what you get and then you have to evaluate your results based on these facts". Student S1 reacted to this statement by expressing that the rule about the number of samples taken needs to be revised: "how many samples are you supposed to take given that you don't know if there are any sub-populations missing". In summary, Group A was concerned with how many samples to take in order to catch the low frequency colors ("the sub-populations"), but also accepted the results as "you get what you get", which to some extent minimalized their concern.

A second example, identified in the work of group B, illustrates the importance for the group that their samples mirrored the distribution of colors in the population. The three students in Group B (S3, S4 and S5) set out to take one sample of four beads from the seventh container. The students poured the beads over the sampling device, but since it proved difficult for them to capture four beads (repeated tries resulted in three, two or zero beads ending up in the sampling device), they decided to collect five samples from the container with replacement and getting varying numbers of beads in each sample. They recorded which colors occurred each time they collected a sample. After having collected five samples, student S3 said: "in our samples we don't see the yellow bead". Student S4 responded by arguing: "that is not totally unreasonable… and in our samples we get the greatest number of red beads, a majority, followed by green and orange". This argument introduced a new element into their model: an ordering of the frequencies of the colors. Even though S4 just argued that their sampling model is not unreasonable, S3 suggested: "Let us take five more samples and see if we can collect some more orange [beads] and the yellow one". They increased the total number of samples to ten. In summary, the students in Group B developed a rule about taking a sufficient number of samples to ensure that (a) low frequency colors would be captured and (b) that the ordering of the frequencies of colors in their samples mirrored the frequencies they observed in the population. The group conjectured that "It feels reasonable that we need to collect a larger sample if the population is larger", indicating that sample size is proportional to the population. This resulted in the group adjusting the sample size. The final rule the students applied in their sampling model was to take ten samples of size five, replacing each sample before taking a new sample.

In the third example, we show how Group C tackled their concern with capturing the low frequency beads by making sure that the number of samples is adequately high. Group C decided to initially use the rule to take a set of two samples, each containing ten beads, placing back the sampled beads from the first sample back into the population before taking the second sample. In other words, they sampled with replacement. However, after taking the first set of two samples of sample size ten with replacement from the seventh container, they observed that the population contained a single yellow bead which was not captured in their two samples. Group C decided to continue to collect beads according to their model (sets of two samples of size ten, with replacement) to see how many sets of two samples were needed to capture the single yellow bead. The results of their exploration are displayed in Table 2.

Table 2. Sampling results by Group C to assure capturing the yellow bead

Set 1		**Set 2**		**Set 3**		**Set 4**	
Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8
3 red	6 red	4 red	3 red	4 red	4 red	3 red	5 red
4 orange	2 orange	3 orange	3 orange	3 orange	3 orange	3 orange	2 orange
3 green	2 green	3 green	4 green	3 green	3 green	3 green	3 green
-	-	-	-	-	-	**1 yellow**	-

Group C needed four sets to capture the yellow bead. One of the students commented that this was "not totally unexpected, but since two samples were too few, I guess we better increase and take some more [samples] throughout". The fact that four sets of two samples were needed to catch the yellow bead in the seventh container, rather than one set as suggested in their initial sampling model, led Group C to adjust their rule dictating the number of samples to take. The group adjusted their developing sampling model by introducing a multiplicative factor of four in the rule specifying the number of samples to collect, since four times as many samples than specified in their initial rule were required to capture the yellow bead in the seventh container. Group C's rules concerning sample size will be described further in a subsequent result section.

These three examples show that when the students' initial models did not result in catching the low frequency colored (yellow) bead, the students considered whether or not their initial sampling models were reasonable. In all three examples the students developed their sampling model by increasing the number of samples taken as a way to deal with their concern. In addition, the examples illustrate three different ways in which the students revised their rules in order to capture low frequency colors in their sampling models. However, the consequences of actually catching the low frequency colors in the larger populations introduced other concerns for the students, which are discussed in the next section.

Not to overestimate the low frequency colors

The second theme identified the concern of the students not to overestimate the low frequency colors in a population. This concern arose primarily after the groups had implemented their sampling models on the six containers and compared their results to the actual distributions of colors given in the envelope in the end of the activity. Evaluating the predictions of their sampling models about the frequency distribution of colors prompted the students to examine the relationships among the number of samples, sample size, and replacement as the basis for making their inferences about a population. We illustrate how students further developed their sampling model as a consequence of this evaluation with the interactions among the students in Group B.

The students in Group B discussed why the number of blue beads was overestimated by their sampling model in the skewed container with the population of 10 000 beads shown on the right in Figure 1. Initially, when comparing the correct distributions for the 20 and 400 bead populations to their inferences, the students described their inferences as "quite good", "nice" and "close" regarding both the number of colors and their frequency distributions. However, these students were not as satisfied with their inference about the distribution of colors in the 10 000 bead container with the skewed distribution (see Table 3).

The students argued that it was not surprising that their sampling model failed to sample either of the two black beads, since "getting two black beads out of 10 000 is not reasonable". However, they expressed concern with their inference of the population containing 200 blue beads, arguing that this was "not good enough" given that the actual number of blue beads was only ten.

Table 3. Inferred distribution of colored beads in the 10 000 container with a skewed distribution found by Group B

Color	Actual Frequency	Inferred Frequency
Dark pink	5821	6200
Green	3179	2800
Red	700	600
Orange	200	200
Yellow	88	0
Blue	10	200
Black	2	0

Referring to the rule of their sampling model to use ten samples of size five with replacement, student S5 argued that “we got one blue bead in our sample which resulted in [a prediction of] 200 blue beads in the population since we collected too few beads [in our sample]”. Student S3 followed up on this statement and made some calculations:

> If we had collected 20 beads [instead of 5 beads] ten times we would have had collected a total of 200 beads. One collected blue bead under those circumstances corresponds to 0,5% [of the whole population]. Scaling up this blue bead then results in 50 blue beads which would had been more reasonable.

Student S3 argued that an increase in sample size from five beads to 20 beads would have resulted in an inference of 50 blue beads instead of 200. According to S3 this would have been more “reasonable” in comparison to the actual ten blue beads in the population. When the students saw the discrepancy between the inference made based on their sampling model and the actual distribution, this triggered them to reconsider the relationship between the size of the sample and the inferences about the distribution of low frequency colors. The group concluded that collecting a larger sample when the population is larger will reduce the risk of making an inference that overestimates the low frequency colors.

The example shows that when the students compared their results to the known distribution, they started to re-evaluate their sampling model and argued for using larger sample sizes with larger populations. Group B was happy with taking ten repeated samples of size five for the smaller populations, but justified their reasoning for taking larger samples for larger populations by offering calculations of why their sampling model was not good enough for the 10 000 bead population. Nevertheless, the students did not revise their sampling model in the written work that the group handed in, possibly because they already had formulated and written an explicit formula as an answer.

Formalizing the relationships facilitating inferences

In our analysis, we found that the students assumed that they were expected to find some type of formula at the end of the activity. Utterances such as “We first need to find a system that works for the other populations, so we need to find a formula to formulate a reasonable hypothesis” were found in several groups. All groups presented the rules and relationships in their final models about generating the samples in order to make inferences about the populations as either strict mathematical formulas or pseudo-mathematical formulas mixing natural language and mathematical notation. We found that the students' sampling models could be characterized in terms of the rules developed by the students regarding (1) sample size, (2) the number of samples, and (3) replacement or no replacement.

We found that all groups used the same fundamental strategy to make an inference about the distribution of colored beads in a population of *N* beads. Namely, students made a statistical inference about the distribution of colors based on scaling up the proportions found through their sampling model to the whole population. Further, six of the sampling models analyzed included rules to collect more than one sample (resampling) and all of them applied the rule of replacement: that is, putting a collected sample back into the population before taking the next sample.

The rules determining the sample size and the number of samples collected in the sampling models were categorized as either fixed or dynamic. A fixed sample size or a fixed number of samples is a constant value that is independent of the size of the population. A dynamic sample size or number of samples varies depending on the situation at hand. This variation could, for example, take the form of an explicit functional dependence on the population size, or the variation could be a dynamic rule dictating different actions affecting the number of samples depending on what the previous sample revealed. The results of the categorization of the sample sizes and the number of samples used in the eight groups' sampling models are shown in Table 4.

Table 4. The groups' sampling models characterized with respect to dynamic and fixed sample size and number of samples

		Number of samples	
		Dynamic	Fixed
Sample size	Dynamic	2	4
	Fixed	1	1

Table 4 shows that most groups' sampling models (four out of eight) used a fixed number of samples with a dynamic sample size as the basis for their inferences about a population. Two groups of students used dynamic rules for determining both the sample size and the number of samples. Finally, two groups used fixed samples sizes. One of the groups used a fixed sample size in combination with a dynamic rule for determining the number of samples; whereas the other group used a fixed rule for the number of samples. In the following four sections, we illustrate the various types of sampling models the groups developed in each category and elaborate on why and how the groups developed their models.

Dynamic sample size and fixed number of samples

Four groups developed sampling models with rules where the number of samples collected is fixed (independent of the population size) but where the sample size is dynamic (dependent of the population size). One illustrative example in this category is Group D's sampling model. The students in Group D began with the premise to only take a single sample since "that it is what is being done in real-life polling situations". Group D's rule for determining the sample size originated early in their discussion when the group concurred that they wanted to take one sample with a size dependent on the population size. The group set out to find an increasing function with a decreasing rate and used a calculator to explore different functions (for example $\log_{10} N$, $(\log_{10} N)^2$, $\log_2 N$ and $\ln N$). After some trial and error, they decided to use a sample size that was equal to the square root of the population, $\sqrt{N}$. When discussing the features of the formula the students realized that $\sqrt{N}$ produced too small sample sizes for small populations. The students found this disconcerting but resolved the situation by modifying their rule to take a sample of $\sqrt{N} + 5$ beads. The students argued that the additional 5 beads in the sample did not make a big difference for the sample size for large population, but that it made a big different for small populations, to be able to make a reasonable inference about the population.

Dynamic sample size and dynamic number of samples

Two of the eight groups developed a sampling model in which the rules for determining the number of samples and the sample size to base their inferences on both depended on the population size. Group C developed a breaking point rule for sample size depending on the population size as follows: "If the population size is less than 20, then use a sample size of 10. If the population size is larger than or equal to 20, then use a sample size of 20 beads". The students in Group C had found it convenient to use the sampling device which sampled about 20 beads. This group decided that 20 was a suitable sample size for larger populations. However, the rule of Group C shows a smaller sample size if the population is less than 20. These students wanted to use the seventh container to explore their sampling model realistically, which to the students meant not using a sample size that equaled the size of the whole population. This exploration turned out to have consequences for how their rule specified the number of samples.

Group C's initial rule for how many samples to take when the population is greater than 20 was: "Take *(size of population)/20* (rounded to nearest integer) samples of 20 beads with replacement". The argument the students in Group C had for using a rule with a variable number of samples was to increase the number of samples proportionally to the population size to get a large enough set of samples to base their inferences on. However, when validating their initial rule using the seventh container (as discussed above, see Table 2), they were led to modify their rule by quadrupling the number of samples: "Take *((size of population)/20) · 4* (rounded to nearest integer) samples of 20 beads with replacement". In other words, Group C modified their initial dynamic rule to a dynamic rule specifying to take four times as many samples as first stipulated.

Fixed sample size and dynamic number of samples

One group used a fixed sample size and dynamic number of samples. Group E's sampling model used a fixed sample size of ten beads. They used a dynamic rule for determining the number of samples (with replacement) to be four, eight and 16 samples for the 20, 400 and 10 000-populations respectively. Based on how one realistically would collect samples in a general situation, the group argued that a fix sample size of ten was a feasible size, but that the number of samples needed to be larger for larger populations. The group discussed various alternatives for the number of samples for the three given populations, and finally agreed on four, eight and 16 samples out of esthetic reasons: "…yeah, having powers of twos is really neat and kind of beautiful".

Fixed sample size and fixed number of samples

Only one of the groups' sampling models developed rules for the sample size and the number of samples to be fixed and thus independent of the population size. Group B's initial rule was to use one sample of size four. However, after having difficulties with actually catching the beads in the sampling device, they realized that they (a) needed to take multiple samples and (b) needed to find a way to make sure each sample caught the same number of beads. After a brief discussion, they decided to use a fixed sample size of five beads and to take a fixed number of samples (ten samples) with replacement regardless of the size of the population.

Discussion and conclusions

The students' developing sampling models examined in this paper highlight the complexities of the ideas the students struggle with in making inferences about a population with respect to a categorical variable with nominal values. Since this topic has not been well researched, either from a theoretical statistical or an educational point of view, the use of the models and modeling perspective facilitated an approach to gain new insights into students' reasoning using the design of a sense-making activity for students using hands-on materials that elicited their ideas. In addition, the models and modeling perspective helped make the nuances of the students' reasoning visible in terms of students' rules for taking samples, their operations on and properties of collected samples, and their ideas about the relationships between the collected samples and a given population.

The two first themes characterizing the relationships between the samples and a population used by the students (catching the low frequency colors and not overestimating the low frequency colors) might in part be a consequence of the design of the task the students worked on. The task included the seventh container with its distribution (nine red, five green, five orange, one yellow bead) and the first question in the task asked about the number of different colors in a container. The data showed how the single yellow bead influenced the students' thinking and their resulting sampling models by engaging in self-evaluating and revising their developing models (Ärlebäck & Doerr, 2014; Lesh & Doerr, 2003). Experimenting and testing their developing sampling models on the seventh container led the students to revise their initial models by adjusting the sample size, the number of samples, or how to manipulate the collected samples to be able to make reasonable inferences about the population. In terms of sample representativeness and sample variability (Ben-Zvi et al., 2015), the students' strongly expressed concern to catch all the colors – and especially the yellow bead – showed that students' understanding of sample representativeness was given primacy over sample variability. At the same time, we saw that Group B's overestimation of a low frequency color in a large population led the students to adjust their sampling model to counter such unreasonable inferences. This suggests that a recognition of the role of sample variability is emerging in the students' thinking. Indeed, one way the students attempted to incorporate a balance between sample representativeness and sample variability (Saldanha & Thompson, 2002; Shaughnessy, 2007) in their developing models was by introducing rules for multiple samples in their sampling models. Using multiple samples in their models addressed the concern of the students to catch all low frequency colors since the more samples that were collected increased the likelihood of capturing all colors. In addition, examining the collection of samples provided a way to not overestimate the low frequency colors when scaling up proportionally to the whole population.

Three aspects of the task design should be noted. First, the two questions in the task, intended to elicit students' ideas, are very different in nature. The first question asks for an inference about a single variable (the number of colors), and the second question asks for the entire distribution of all the values the categorical variable can take in the population. Based on our data and analysis, it appears that the fact that the students were presented with, and hence considered, both questions simultaneously affected the students' thinking and the development of their sampling models. Hence, it is possible that some students' ideas with respect to the two questions were conflated as they developed their models. Second, the data suggest that the sampling device might have unintentionally limited the students' thinking and prompting them to use a sample size of 20 rather than to approach the issue of deciding on the sample size based on their own ideas. Both these aspects may have acted to limit students' autonomy in developing their sampling models. These aspects could be modified in a redesign of the task and examined in future research. Third, the chosen sizes of the given populations (*N* = 20, 400, 10 000) were aimed at promoting the students to think about the potential role of population size in their sampling models. The 20 bead population size might have created ambiguity for some of the students since the two main questions in the task are effortlessly fully determined by sampling all 20 beads. In a redesign of the task, we would suggest to only focus on the second question of the task, to use a more flexible sampling tool, and to use population sizes of *N* = 100, 1 000 and 10 000. Given these adaptations combined with a follow-up whole-class discussion of the students' sampling models, we suggest that the task could be productively used with secondary level students.

One of the conclusions of this study is that the concerns and challenges students have in understanding and balancing sample representativeness and sample variability, documented when working with single value statistics such as mean, median, mode and proportions (Batanero et al., 2020; Ben-Zvi et al., 2015; Lipson, 2003; Saldanha & Thompson, 2002; Shaughnessy, 2007), are also found in the case of categorical variables taking nominal values. Further, our focus on the students' reasoning and developing sampling models extends the work done on interpretation of tables and graphs representing categorical data (Budgett & Puloka, 2019; Casey et al., 2018a; 2018b) by considering both distributions of a categorical variable when the number of values is greater than two. In addition, the present study extends the scope of the prior work focusing on categorical variables by examining approaches to sampling a categorical variable and drawing inferences based on the samples.

Future research

The result of this study provides insights into how the task used could be redesigned to give more options to students to use their own ideas with respect to choosing the sample size to use. A redesign of the sampling device or the provision of multiple sampling devices could more effectively facilitate the sampling procedure, without limiting the students' thinking. In addition, separating the two questions in the task to investigate students' reasoning in each separate situation and what sampling models they then develop appears to be a logical next step to take. Given that the two questions are so different in nature, it might be productive to investigate the students' developing sampling models when they only are presented with one of the questions, to see how they deal with sampling categorical data when a single statistic is sought such as number of values or the proportion of values or when the whole distribution of a given attribute of a categorical variable is the achievable objective. In particular, it could be interesting to look further at students' reasoning when they draw inferences on categorical values, especially in connection to the students' use of bootstrapping ideas (Engel, 2010; Hesterberg, 2006; McLean, 2015) since ideas about multiple consecutive samples emerged in the work of the students.

References

Ärlebäck, J. B., & Doerr, H. M. (2014). Preserving students' independence by encouraging students' self-evaluation. In H. Silfverberg, T. Kärki, & M. Hannula (Eds.), *Nordic research in mathematics education – Proceedings of NORMA14* (pp. 257–266). Turku, Finland: The Finnish Research Association for Subject Didactics.

Barbosa, J. C. (2006). Mathematical modelling in classroom: A socio-critical and discursive perspective. *ZDM*, *38*(3), 293-301. https://doi.org/10.1007/BF02652812

Batanero, C., Begué, N., Borovcnik, M., & Gea, M. M. (2020). Ways which high-school students understand the sampling distribution for proportions. *Statistics Education Research Journal*, *19*(3), 32-52. https://doi.org/10.52041/serj.v19i3.55

Batanero, C., Estepa, A., Godino, J. D., & Green, D. R. (1996). Intuitive strategies and preconceptions about association in contingency tables. *Journal for Research in Mathematics Education, 27*(2), 151–169. https://doi.org/10.5951/jresematheduc.27.2.0151

Ben-Zvi, D., Bakker, A., & Makar, K. (2015). Learning to reason from samples. *Educational Studies in Mathematics*, *88*(3), 291-303. https://doi.org/10.1007/s10649-015-9593-3

Böcherer-Linder, K., Eichler, A., & Vogel, M. (2016). The impact of visualization on understanding conditional probabilities. *Proceedings of the 13th International Congress on Mathematical Education*. Hamburg, Germany. Retrieved from https://iase-web.org/documents/papers/icme13/ICME13_S1_Boechererlinder.pdf

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*, 77–101. https://doi.org/10.1191/1478088706qp063oa

Budgett, S., & Puloka, M. (2019). Making sense of categorical data – question confusion. In S. Budgett (Ed.), *Decision making based on data. Proceedings of the satellite conference of the International Association for Statistical Education (IASE)*. Kuala Lumpur, Malasya. Retrieved from http://iase-web.org/documents/papers/sat2019/IASE2019%20Satellite%20114_BUDGETT.pdf

Casey, S., Albert, J., & Ross, A. (2018a). Developing knowledge for teaching graphing of bivariate categorical data. *Journal of Statistics Education*, *26*(3), 197-213. https://doi.org/10.1080/10691898.2018.1540915

Casey, S., Hudson, R., & Ridley, L. (2018b). Students' reasoning about associations of categorical variables. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10)*. Kyoto, Japan. Retrieved from https://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_2E2.pdf?1531364243

Dodge, Y. (2008). *The concise encyclopedia of statistics*. Springer. https://doi.org/10.1111/j.1751-5823.2008.00062_25.x

Doerr, H. M., delMas, R., & Makar, K. (2017). A modeling approach to development of students' informal inferential reasoning. *Statistics Education Research Journal*, *16*(2), 86-115. Retrieved from https://iase-web.org/documents/SERJ/SERJ16(2)_Doerr.pdf

Doerr, H. M., & English, L. D. (2003). A modeling perspective on students' mathematical reasoning about data. *Journal for Research in Mathematics Education, 34*(2), 110-136. https://doi.org/10.2307/30034902

Engel, J. (2010). On teaching bootstrap confidence intervals. In C. Reading (Ed.), *Proceedings of the Eighth International Conference on Teaching Statistics*. Ljubljana, Slovenia. International Statistical Institute. Retrieved from http://iase-web.org/documents/papers/icots8/ICOTS8_4B2_ENGEL.pdf?1402524970

Everitt, B. S., & Skrondal, A. (2010). *The Cambridge dictionary of statistics.* Cambridge University Press. https://doi.org/10.1111/j.1751-5823.2011.00149_2.x

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report: A pre-K–12 curriculum framework*. American Statistical Association. Retrieved from https://www.amstat.org/asa/files/pdfs/GAISE/GAISEIIPreK-12_Full.pdf

Hestenes, D. (2010). Modeling theory for math and science education. In R. Lesh, P. L. Galbraith, C. R. Haines, & A. Hurford (Eds.), *Modeling students' mathematical modeling competencies: ICTMA 13* (pp. 13-41). Springer. http://doi.org/10.1007/978-1-4419-0561-1_3

Hesterberg, T. (2006). Bootstrapping students' understanding of statistical concepts. In G. Burrill (Ed.), *Thinking and reasoning with data and chance. Sixty-eighth National Council of Teachers of Mathematics Yearbook* (pp. 391-416). Reston, VA, USA: National Council of Teachers of Mathematics.

Lehrer, R., & Schauble, L. (2010). What kind of explanation is a model? In M. K. Stein & L. Kucan (Eds.), *Instructional explanation in the disciplines* (pp. 9-22). Springer. http://doi.org/10.1007/978-1-4419-0594-9_2

Lesh, R. A., & Doerr, H. M. (Eds.). (2003). *Beyond constructivism: Models and modeling perspectives on mathematics problem solving, learning, and teaching*. Lawrence Erlbaum Associates.

Lipson, K. (2003). The role of the sampling distribution in understanding statistical inference. *Mathe matics Education Research Journal*, *15*(3), 270-287. https://doi.org/10.1007/BF03217383

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). *Big data: The next frontier for innovation, competition and productivity*. McKinsey Global Institute.

McLean, J. A. (2015). *Eliciting bootstrapping: The development of introductory statistics students' informal inferential reasoning via resampling*. Dissertation. https://surface.syr.edu/etd/396

Organisation for Economic Co-operation and Development (OECD). (2013). *PISA 2012 assessment and analytical framework. Mathematics, reading, science, problem solving and financial literacy*. OECD. https://doi.org/10.1787/19963777

Saldanha, L., & Thompson, P. (2002). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics, 51*(3), 257-270. https://doi.org/10.1023/A:1023692604014

Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 957–1009). Information Age Publishing.

Watson, J. M., & Callingham, R. (2014). Two-way tables: Issues at the heart of statistics and probability for students and teachers. *Mathematical Thinking and Learning, 16*(4), 254–284. https://doi.org/10.1080/10986065.2014.953019