Growing Certain: Students' Mechanistic Reasoning about the Empirical Law of Large Numbers

A Dissertation

SUBMITTED TO THE FACULTY OF

UNIVERSITY OF MINNESOTA

BY

Ethan C. Brown

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Robert delMas, Adviser

Andrew Zieffler, Co-Adviser

May 2019

Acknowledgements

This dissertation represents the work of many people, visible and invisibly. Most obvious is the incredible community of support that I found in the Statistics Education track, and more generally in the Quantitative Methods in Education program. Joan Garfield, at the urging of Michael Harwell, followed her vision and created this program, the first of its kind in the United States and the second in the world, and provided valuable feedback and direction to me in my first years in the program. Michelle Everson welcomed me to the program as my first advisor and taught me so much about teaching statistics. A warm community of students welcomed me into the program, and a swirl of collaborations, debates, and laughter swirled around those who joined around the same time—Nicola Justice with our long conversations about statistics theory, Liz Fry with Frying Hash Browns, and Anelise Sabbag with X-Men and heavy metal—and all the other members of the stat ed family.

Bob delMas has been at the center of this all, as consistently helpful, thoughtful, and engaged an advisor as I could ever imagine. Bob is the kind of person who hosted my parents for my Masters commencement and programmed an application on Thanksgiving to explore ideas related to my thesis. Andy Zieffler has been a supportive co-advisor and generous collaborator on our many projects together during my time in graduate school, especially the QME package. Michèle Mazzocco has been an amazing mentor and I grew up so much in her lab and in the AC/DC room. Keisha Varma provided valuable perspectives from science education on both of my exam committees. I'd like to thank the Ed Psych support staff who kept everything running: the infinite patience and helpfulness of Lori Boucher, the warm liveliness and engagement of Sharon Sawyer, and many other staff members who’ve been so helpful.

The story of my time in graduate school has been very tied as well to the amazing international community at Minnesota Graduate Club. Thank you to To Nga Hoang, John Ssenkusu, Samadangla Ao, and to many others who helped me grow so much as a person, and to prepare me for the biggest adventure of my life—my wild friendship, romance, and spiritual partnership with Ha, who has supported me on this journey by believing in me, filling in domestic logistics to make things work, providing a brave and truthful example, giving gut-splitting laughter and delight, and regularly kicking my butt to keep me on track. In broader life, there are many others who have my dearest gratitude but this dissertation is already a bit on the long side, so I’ll close by thanking my old dear friends—Katie, Lexi, Lucian, Chris, Adam, Mel, Robin, and Erica—and mom, dad, brother, sister-in-law, and nephew for inspiring me and supporting me when I needed it most.

Finally, I’d like to acknowledge that not all the support and privilege given to me was freely and voluntarily given. This dissertation was written on Daḳota land and I and my ancestors have gained material benefit from racist and unjust policies and practices. I pray that I can learn to use these privileges for liberation.

Abstract

Extensive research has documented students' difficulty understanding and applying the Empirical Law of Large Numbers, the statistical principle that larger random samples result in more precise estimation. However, existing interventions appear to have had limited success, perhaps because they merely demonstrate the Empirical Law of Large Numbers rather than support students' conceptual understanding of *why* this phenomenon occurs. This dissertation developed a sequence of activities, *Growing Certain,* which provided support for two mechanistic explanations of the Empirical Law of Large Numbers for students in a simulation-based introductory statistics course: *swamping*, the decreasing influence of extreme values on the mean as sample size increases, and *heaping*, the increasing concentration of possible sample means around the population mean. Five students participated in over six hours of one-on-one clinical interviews, with analysis focused on one focal participant, "S". S's responses were analyzed using a detailed coding of S's articulation of mechanism components. S already displayed strong inclination towards swamping in the pre-interview questions, and their articulation of swamping became more sophisticated as they progressed in Growing Certain. However, S's understanding of the connections between population and sample were weak throughout, and S had a lot of difficulty reasoning about multiple sample means simultaneously in a sampling distribution. S's lack of abstraction of the sample mean appeared to support them in attending to the dynamics of swamping, but hindered them in being able to reason about heaping. Future research could examine representations that bridge swamping and heaping, and to examine individual differences in attention to the mechanistic components of the Empirical Law of Large Numbers.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Making data-based decisions under uncertainty is no longer a skill only important for scientists and policy-makers, but is necessary for all democratic citizens. Particularly important for making such decisions is the Empirical Law of Large Numbers: the statistical principle that large random samples increase the precision of estimation (Fiedler, Walther, & Nickel, 1999; Utts, 2003). For example, a poll of 10 people is likely to provide a much less precise estimate than a poll of 1000 people. Many studies, building on the classic work of Kahneman and Tversky (1972), have shown that people frequently ignore or reason inconsistently about the effect of sample size depending on the context (see Lem, Van Dooren, Gillard, & Verschaffel, 2011). Despite extensive study of sample size reasoning, there is no cogent theoretical framework of why, how, and when people reason about sample size (Lem et al., 2011). Some of the problem appears to be confusion about the difference between the distribution of individual values and the hypothetical distribution of the means of many samples (Sedlmeier & Gigerenzer, 1997), but confusions and inconsistencies are present even when people appear to have mastered this distinction (delMas, Garfield, & Chance, 2006; Well, Pollatsek, & Boyce, 1990). As statistics becomes more essential in a data-rich society, and as education standards increasingly emphasize the importance of statistics education for everyone (Common Core State Standards Initiative, 2010), it is imperative that we understand how students think about statistical issues such as the Empirical Law of Large Numbers and give them opportunities to understand why such principles apply.

1.1 The Empirical Law of Large Numbers in contemporary statistics education

Sample size is fundamental in statistics almost by definition, since the field is devoted to understanding the behavior of *aggregations* of the data. Moreover, the effect of sample size is what drives the convergence of data-based evidence towards a true probability or expected value, and is thus fundamental to statistical power as well. This is often referred to as the Law of Large Numbers in both psychology (Nisbett et al., 1983) and statistics education (Pratt et al., 2008). As Sedlmeier and Gigerenzer (1997) point out, however, the Law of Large Numbers is a statement about the *limit* of deviations from the mean as *n* approaches infinity. It does not say anything about finite samples, or trends as sample sizes increase, and is therefore not applicable to any of the sample size tasks reviewed here (the Central Limit Theorem is inapplicable for similar reasons). Therefore this review adopts their term, the *Empirical* Law of Large Numbers (Freudenthal, 1972; Sedlmeier & Gigerenzer, 1997) to denote the widespread and fairly general phenomenon that larger samples are more likely to provide sample means/proportions that are close to the true mean/proportion.

An educational reform movement has made many contributions to statistics education, and standards documents, new assessments, and new curricula are intended to focus the undergraduate course on meaningful conceptual understanding rather than on rote formulas and procedures (GAISE College Report ASA Revision Committee, 2016). Using simulation-based methods to help students understand inference has been posited as one manner to help students build conceptual understanding of the logic of hypothesis tests. Instead of solving formulas based on the normal distribution, perhaps after sitting through an abstract proof of the central limit theorem, students actually generate many random samples (or random assignments) themselves based on the problem setup, and directly view and evaluate the empirical sampling distribution (Cobb, 2007). This approach has been found to have several potential benefits for teaching the conceptual understanding of inference (e.g., Tintle, Topliff, Vanderstoep, Holmes, & Swanson, 2012).

However, do students in simulation-based courses really understand the behavior and dynamics of empirical sampling distributions? The dominant software representations that are used to explore empirical sampling distributions, including code-based interfaces in R (R Core Team, 2018) as well as approaches using specialized educational software such as TinkerPlots™ (Konold & Miller, 2017), Fathom™ (Finzer, Erickson, & Binker, 2005), StatKey (Morgan, Lock, Lock, Lock, & Lock, 2014), the RossmanChance applets (http://rossmanchance.com/applets), or Sampling SIM (delMas, 2002), do not make clear how sample size affects the distribution of sample means. Instead, the effect of sample size is a contextual factor that students seem to struggle to integrate (Brown, 2015; Chance, delMas, & Garfield, 2004). Simulation-based inference approaches may clarify the logic of hypothesis testing, but these approaches do not make clear how the aggregation of growing samples at certain sizes produces distributions with certain sizes and shapes.

This dissertation provides one of the first in-depth qualitative studies of conceptual understanding of the Empirical Law of Large Numbers for students in a simulation-based inference course. The extensive literature on people's understanding of the effect of sample size is re-examined in order to identify the apparent conceptual understandings that have led to such a variety of performance on sample size tasks. These conceptual understandings are then linked to recent research tregarding the role of *causal* understanding in conceptual change. The study itself builds on the knowledge students are learning in a simulation-based inference course, insights and tasks from prior research on the Empirical Law of Large Numbers, and prior frameworks for characterizing causal reasoning about scientific phenomena.

1.2 Description of the study

The participants were five students in a tertiary general-education introductory statistics course during Spring 2018 which used the simulation- and randomization-based CATALST curriculum (Garfield, delMas, & Zieffler, 2012; Zieffler & Catalysts for Change, 2017), and who had already regularly created probability simulations and randomization-based hypothesis tests in TinkerPlots™. These five students participated in five video-recorded clinical interviews while performing a sequence of tasks, called Growing Certain, that provided increasingly rich conceptual support for causal thinking about sample size. Because the literature review revealed the potential importance of mechanistic causal reasoning in conceptual change, and the relative lack of support for causal reasoning in prior interventions targeting the Empirical Law of Large Numbers, this inspired two exploratory research questions, inspired by Russ, Scherr, Hammer, and Mikeska (2008) and Parnafes & diSessa (2013):

1. *How and when are students reasoning mechanistically before, during, and after Growing Certain? How does each element of Growing Certain (prompts, representations, etc.) interact with students' mechanistic reasoning about the Empirical Law of Large Numbers?*
2. *How can we describe conceptual change when it happens in the context of Growing Certain? How do the changes observed relate to the task design, including the representations used, prompts, and social interactions?*

Initially, students' thinking about the effect of sample size was explored in various classic sample-size tasks, including the hospital problem (Kahneman & Tversky, 1972). Students were regularly prompted to describe why they thought certain phenomena would be observed with larger sample sizes, but no technology was provided. Then, students were given activities of increasing complexity to visualize the causal role of sample size for long-run stability. Students first explored the behavior of a sample proportion or mean as sample size grew. The main conceptual target was *swamping* (Well et al., 1990), the idea that a single extreme value influences the mean of a large sample much less than the mean of a small sample. Students visually created random models that incrementally added values onto a sample and were given opportunities to witness how the mean, initially quite unstable, began to settle down near an expected value in a plot of the mean against the sample size. Students were prompted to experiment with trying to change a TinkerPlots™

model of a Cat Factory (Konold, Harradine, & Kazak, 2007) in ways that would change the plot of the mean against the sample size. Students then explored “what if” scenarios of how a slot machine might actually be working, with varying variability and closeness of the true mean to a hypothesized mean, and evaluated what their sample size would need to be to test the hypothesized mean.

A more comprehensive explanation of the Empirical Law of Large Numbers requires understanding why the empirical sampling distribution of means shows an increasing concentration around the true mean value as the sample size increases, a phenomenon here termed *heaping*. In activities inspired by the Blink Game (Konold & Kazak, 2008), students first explored this phenomenon by growing multiple samples simultaneously and plotting the means. They then explored the underlying mechanism by examining the *unique* outcomes that arise in simulations, and being given opportunities to note that there are more possible outcomes near the mean. Finally, students built up a physical model of the theoretical sampling distribution using construction blocks, and were asked about the connections of the physical model to the simulations they had previously created. The sessions concluded with a post-interview, where students explored new sample size problems similar to those in the first interview. Students then did the same problems as the first interview, and compared and contrasted their new answers with their original answers. Finally, students were probed about the connections between the dynamic TinkerPlots™ models and representations and the sample size problems, and asked generally about how they would explain how sample size causes long run stability.

All students were videotaped and had their computer screens recorded. One student of the five, the focal student, was selected for detailed coding and analysis. The focal student's responses were analyzed using a coding system inspired by a combination of mechanistic discourse analysis (Russ, Scherr, Hammer, & Mikeska, 2008) and microgenetic learning analysis (Parnafes & diSessa, 2013; Siegler, 2007). This coding attempted to capture the details of the student's mechanistic reasoning, and to examine how that reasoning changed and evolved over the course of the study in order to answer the two research questions.

1.3 Structure of the dissertation

Chapter 2 presents a literature review on understanding and learning the Empirical Law of Large Numbers. The chapter distills eight student conceptions from prior research regarding the effect of sample size. The chapter next examines prior interventions intended to teach the Empirical Law of Large Numbers, discusses the limitations of prior findings, and argues from recent conceptual change research that a mechanistic causal explanation of the Empirical Law of Large Numbers may be more effective. The chapter concludes by arguing that swamping and heaping may be most promising for supporting a mechanistic causal explanation.

Chapter 3 presents the methodology for the present study. The chapter begins by outlining the targeted causal mechanisms for sampling variability and how Growing Certain supports them. The recruitment of participating introductory statistics students is then described, including further detail about the CATALST curriculum (Garfield et al., 2012; Zieffler & Catalysts for Change, 2017). The Growing Certain activity sequence is then described in detail, including the changes that it underwent during pilot testing, followed by comments about how the interviews were administered. Finally, the chapter describes the initial qualitative analysis procedure based on mechanistic coding and microgenetic analysis.

Chapter 4 presents the results of the qualitative analysis for the focal student. First, a summary of how the coding system evolved during the actual analysis is presented, along with overall summaries of the codes. Each interview is then summarized in a separate section, with separate subsections for the distinct segments of those interviews. Each interview section opens with a summary of the codes that occurred in that interview, and each subsection opens with a detailed look at the mechanistic elements and relationships between elements that were coded in that interview segment.

Chapter 5 discusses the results and relates them back to the literature reviewed in Chapter 2. First, the focal student's reasoning is examined for how it relates to each of the types of mechanistic reasoning identified in Russ et al. (2008). Then, overall comments are offered regarding how mechanistic reasoning about sampling variability can be supported. The chapter then comments on the utility and limitations of the mechanistic coding scheme for understanding student reasoning, and provides broader commentary on teaching implications for statistics education beyond sampling variability. The dissertation concludes by discussing limitations of the study and opportunities for future research.

Appendices include full versions of the tasks and the transcripts of the interviews with the focal student.

Chapter 2

Literature Review

Over 40 years of research has been done into peoples' difficulty understanding and applying the Empirical Law of Large Numbers (e.g., Kahneman & Tversky, 1972). Prior reviews have focused on resolving inconsistencies in performance across different tasks and participants (Lem et al., 2011; Pollard & Evans, 1983; Sedlmeier & Gigerenzer, 1997). Here, the literature on understanding the effect of sample size is first distilled into eight key conceptions that have been revealed through this research to understand *how* people informally reason about, and explain, the role of sample size and variability both correctly and incorrectly. Particularly worthy of note are three types of correct reasoning: *the size-confidence intuition* (larger samples are more representative; Sedlmeier, 1999), *swamping* (means of larger samples are less influenced by extreme values; Well et al., 1990), and *balancing* (extreme values are more likely to balance out in the means of larger samples; *ibid.*).

Secondly, literature on conceptual change is reviewed. How can people change incorrect reasoning to correct reasoning? The Knowledge Revisions Components (KReC) Framework (Kendeou & O'Brien, 2014), drawn from experiments on reading comprehension and refutation texts (e.g., Kendeou, Smith, & O'Brien, 2013), suggests that activating old and new information and strengthening the new information with *causal* explanations is an effective way of revising knowledge. It is argued that to be adequate for understanding sampling variability, an explanation should not only be causal—describing that a cause-and-effect relationship exists—but also *mechanistic*—describing *how* the underlying structure (Russ et al., 2008) of samples causes a given degree of sampling variability. Prior interventions have had some helpful features and successful outcomes, but overall success was mixed and no interventions were found that provided mechanistic explanations of the effect of sample size on sampling variability.

The three correct types of reasoning about sampling variability are then examined in more depth to determine their potential for generating effective inter-level mechanistic explanations. The size-confidence intuition appears difficult to support directly, since students have had difficulty coordinating local variability with global stability (Pratt, Johnston-Wilder, Ainley, & Mason, 2008), and existing interventions' focus on macro-level patterns may not be addressing students' misconceptions about how sampling variability emerges (cf. Chi, Roscoe, Slotta, Roy, & Chase, 2012). Ideas of swamping and balancing may be more promising since understanding the relationship of the values to the mean provides an inter-level causal explanation. However, no previous studies have attempted to support swamping or balancing, and it is unclear how balancing could even be supported.

The review concludes with a theoretical examination of the size-confidence intuition and swamping to examine their adequacy as mechanisms as defined in an influential philosophy of science analysis (Machamer, Darden, & Craver, 2000). The results of the literature review and the theoretical analysis lead to a statement of the research questions for contributing to knowledge about conceptual change in sampling variability and the role of mechanistic reasoning in a new series of tasks and representations.

2.1 Conceptions of Sampling Variability

Cognitive psychology and statistics education researchers have since the 1970s been demonstrating people's difficulties integrating sample size and within-group variability when making intuitive or informal inferences (e.g., Kahneman & Tversky, 1972). However, this literature includes a wide variety of tasks that appear to elicit different results from subjects depending on format, and there is so far no cogent theory that accounts for conflicting findings in the literature (Lem et al., 2011). Developing conceptual understanding of formal inference may help people make more normative choices (Fong, Krantz, & Nisbett, 1986). However, developing conceptual understanding of the place of sample size and within-group variability in formal inference is also quite challenging, as the literature on sampling distributions indicates (Chance et al., 2004).

Many studies have documented tasks where people neglect sample size (e.g., Kahneman & Tversky, 1972) and, relatedly, within-group variability (Obrecht, Chapman, & Gelman, 2007). However, there are also plenty of documented tasks where people do attend to sample size (e.g., Bar-Hillel, 1979) and variability (e.g., Nisbett, Krantz, Jepson, & Kunda, 1983). Furthermore, there are a wide variety of potential ways of thinking about sampling variability that researchers have theorized or that study participants have expressed.

Instead of trying to isolate a single factor to explain the variety of responses to sample size tasks, this section attempts to identify students' conceptions of the effect of sample size based on two bodies of research. One body of research analyzes how people intuitively incorporatesample size, within-group variability, and effect size in statistical inference, nearly always in textual and numerical tasks. The other body of research is from the statistics education literature, and involves examining how visually evaluating empirical and hypothetical sampling distributions can scaffold formal inference.

2.1.1 Population similarity and proportional reasoning

Kahneman and Tversky (1972) posit that people have a *representativeness heuristic*, whereby people tend to judge the probability of a sample outcome by the degree to which it: "(*i*) is similar in essential characteristics to its parent population; and (*ii*) reflects the salient features of the process by which it is generated" (p. 430). The authors do not provide justification as to why these two reasoning criteria should be grouped as a single heuristic. Therefore, this literature review will refer separately to (*i*) *population similarity*, the primary focus of their research, and (*ii*) *apparent randomness*, pertaining to judgements of what looks likely or unlikely under randomness. Kahneman and Tversky argue that participants will see sample size as irrelevant when making judgments about the likelihood of sample outcomes because they will attend only to population similarity—the resemblance of the sample mean or proportion to its population counterpart.

Whatever the value of their theoretical account, Kahneman and Tversky (1972) discovered a striking empirical regularity: Adults neglect sample size in a number of contexts and tasks (see Lem et al., 2011 for a review). Their hospital problem has become a classic demonstration:

> A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50% of all babies are boys. The exact percentage of baby boys, however, varies from day to day. Sometimes it may be higher than 50%, sometimes lower.
>
> For a period of 1 year, each hospital recorded the days on which more than 60% of the babies born were boys. Which hospital do you think recorded more such days?
>
> The larger hospital [*Reversed*]
>
> The smaller hospital [*Correct*]
>
> About the same (i.e., within 5% of each other) [*Equal*] (Kahneman & Tversky, 1972)

Normatively, the smaller hospital will have more variability in the percentage of boys born in any given day and therefore should be expected to have more days with over 60% boys (probability = .15 for a single day, assuming 50% boys) compared with the larger hospital (probability = .07 for a single day). However, only 20% of the participating Israeli high school students (n = 50) chose this *Correct* option, with 54% choosing the *Equal* option—the one which underweights the effect of sample size—and the remaining 26% choosing the *Reversed* option, which has the relationship between sample size and sampling variability backwards (terminology courtesy of Well et al., 1990).

One criticism of the problem is the phrasing of the Equal option. For one thing, it is not clear what the "5%" is referring to. If viewed in percentage terms, a likelihood of 15% for the outcome in the smaller hospital is in fact only 8% more than the 7% likelihood in the larger hospital, which is not much larger than the provided 5%. However, Reagan (1989) found little difference among Harvard undergraduates between those who received the original problem phrased as “About the same (i.e., within 5% of each other)” (*n* = 35, Equal = 49%) as compared with those who received a stronger claim of “Exactly the same” (*n* = 31, Equal = 45%).

Moreover, the finding of sample size neglect on many variants of the hospital problem has been repeatedly replicated across several, although not all, conditions and cover stories (e.g., Lem et al., 2011; Well et al., 1990). Kahneman and Tverksy (1972) support their claims that participants’ judgements are based on population similarity on other more complex and perhaps problematic tasks (Bar-Hillel, 1979; Evans & Dusoir, 1977; Olson, 1976) involving directly estimating sampling distributions of differing sizes. The hospital problem has become most popular with researchers for its relatively clean and still-persistent elicitation of sample size neglect.

The theory of "representativeness" as a general problem-solving heuristic has come under sustained criticism due to the difficulty of specifying in advance what subjects will see as “representative” and what is not (e.g., Bar-Hillel, 1979; Gigerenzer, 1996; Pollard & Evans, 1983; Well et al., 1990), and certainly population similarity alone is not sufficient to describe the extensive variability across formally similar tasks (e.g., Olson, 1976) and outcomes found in later studies of sample size neglect—ranging from 4% correct to 100% correct (Lem et al., 2011).

Population similarity, however, still appears to play an important role in sample size judgments and may well be an overgeneralization of proportional reasoning. Well and colleagues (1990, Experiment 3) asked undergraduate psychology students to choose a rationale for their sample size responses. Of those who chose *Equal,* about half chose the rationale, “The number of men who register each day at a post office is not a factor because the problem deals with averages” (Well et al., 1990, p. 300). These students may have learned throughout their education that the average allows one to compare groups of unequal size, and therefore focused on this aspect of the arithmetic mean rather than the potential for variability. Supporting this possibility, Fischbein and Schnarch (1997) administered the hospital problem and a formally similar coin problem to Israeli students in grades 5, 7, 9, and 11 (n = 20 for each grade) along with 18 undergraduate preservice mathematics teachers. For both problems, the proportion of students responding *Equal* was higher in higher grades (with the exception of college students in the coin problem, of which 46% chose the *Equal* answer compared to 75% of the 11th-graders). The authors surmised that older students had learned and then overgeneralized their understanding of proportionality: The same ratio is found in both sample sizes, and both samples are therefore equivalent (Fischbein & Schnarch, 1997).

Although no additional direct evidence for the theory that population similarity judgements are driven by proportional reasoning was found in the sample size literature, there are traces of this in the literature on student understanding of the mean. Indeed, a common justification for the mean for comparing groups is that it allows comparing samples of different sizes, whereas comparing group sums is only appropriate when the sample sizes are the same (e.g., Gal, 1989). In a study of 88 Australian students’ (Grades 3–9) understanding of comparing two groups, Watson and Moritz (1998) found that proportional reasoning facilitated understanding of how to successfully compare two stacked dotplots representing student test scores (Figure 2.1). Students were asked to compare two teams, Pink and Black, where the Black team actually consisted of a subset of the scores of the Pink team but with a clearly higher mean. A Grade 7 student was unable to think proportionally to compare the two groups: “Because there’s more people in Pink than Black, and if you took some of the others away, they would probably be the same.” In contrast, a Grade 9 student was able to more successfully compare the two groups using the arithmetic mean, and then commented: “Even though they had less people it still averages so you work out the average so it’s still fairer . . . that makes it equal averaging”. The average is the “fair” way to compare two groups with different sample sizes. Watson and Moritz’s Grade 9 students showed greater ability to solve this problem (50%, *n* = 28) than did those in Grades 5–7 (14%, *n* = 37) or Grade 3 (0%, *n* = 23).

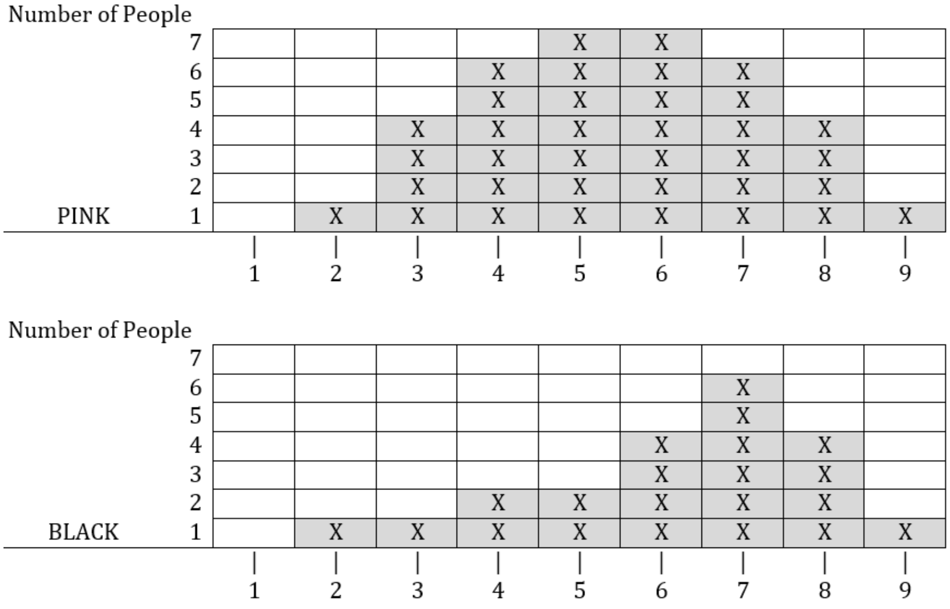


Figure 2.1. Unequal-sized sample comparison recreated from Watson and Moritz (1998). Task derived originally from Gal (1989).

Are students simultaneously learning that means and proportions allow them to fairly compare groups of different sizes—and at the same time that larger and smaller samples would be expected to have similar variability? A participant in Leavy's (2006) study of comparing unequal-sized groups offers her hesitation of comparing using the mean: "But if we have different N's doesn't each one in the smaller set get more significance?" The authors comment that at this point the instructors emphasize how the mean deals with unequal sample size, and that this student is lacking understanding of the mean as a proportional measure. The student still wonders: "I wonder how big does the difference [in sample size] have to be before you can't use the mean to compare them?" Though the instructors are right to teach the mean as a proportional measure, the student is also correct that sample size still makes a difference. Although it is unclear what she was referring to, her statement may hint at the fact that the larger groups' mean is less influenced by each case than is the smaller groups' mean, the phenomenon that Well et al. (1990) call *swamping*.

Overgeneralization of proportional reasoning with samples is closely related to an important feature of all of the sample size neglect tasks in Kahneman and Tversky (1972): All ask statistically naïve participants to make judgements related to sampling distributions, i.e. the probability distributions of means and proportions. Considering the well-documented difficulties students have with sampling distributions (e.g., delMas, Garfield, & Chance, 1999), this may be a source of some of students' difficulties.

2.1.2 Confusing Frequency and Sampling Distributions

Proportional reasoning is necessary to accurately compare sample datasets with unequal sizes, as shown above, but is not sufficient to understand the long-run distribution of means and proportions that emerge upon repeated random sampling. Could neglecting sample size be simply due to not understanding sampling distributions? Evans and Dusoir (1977) were among the first to propose this possibility, though their version of the maternity ward task excluded the *Equal* option and therefore is hard to compare to the original Kahneman and Tversky (1972) findings.

Bar-Hillel (1979) contributed a classic task, the polling problem, that demonstrated sensitivity to sample size when asked about a single sample rather than a sampling distribution:

> Two pollsters are conducting a survey to estimate the proportion of voters who intend to vote YES on a certain referendum. Firm A is surveying a sample of 400 individuals. Firm B is surveying a sample of 1000 individuals. Whose estimate would you be more confident in accepting?
>
> Firm A's [Reversed]
>
> Firm B's *[Correct]*
>
> About the same *[Equal]* (Bar-Hillel, 1979, p. 249)

About 80% of the paid University of Oregon participants chose the *Correct* response, with only 14% choosing *Equal*. This contrasts sharply with the 56% choosing *Equal* in the original hospital problem (Kahneman & Tversky, 1972). The task is, however, substantially different from the hospital problem in several other ways besides for being phrased in terms of single samples. For one thing, the hospital problem asks about the tails of the distribution—*more than 60% boys* (see Well et al., 1990). Also, the polling problem asks about which estimate would inspire more *confidence*. In a study of introductory statistics students who had not received formal instruction about sample size, Brown (2015) found that students' performance on this item was relatively uncorrelated with the hospital problem (Kahneman & Tversky, 1972), the post office problem, and the geology problem (both from Well et al., 1990), which were all rephrased to be about judging which of two samples was "more likely" to be close to the true population value. These results suggest that asking about "confidence" may be different than asking about likelihood, since awareness of mathematical validity and a personal choice may diverge (Amsel et al., 2008). Additionally, some argue that there is no way to evaluate if there are fundamental psychological differences between subjective confidence and probabilities, making them incommensurate (e.g., Hullman, Resnick, & Adar, 2015).

Nevertheless, researchers since Bar-Hillel (1979) have provided many other demonstrations of sensitivity to sample size in similar contexts that were phrased in more probabilistic language. In their investigation of undergraduate psychology students' understanding of sample size and the variability of the mean, Well and colleagues (1990) found students performed well on their post office problem (apparently inspired by the medical survey problem in Kahneman & Tversky, 1972):

> When they turn 18, American males must register for the draft at a local post office. In addition to other information, the height of each male is obtained. The national average height of 18-year-old males is 5 feet, 9 inches.
>
> Yesterday, 25 men registered at post office A and 100 men registered at post office B. At the end of the day, a clerk at each post office computed and recorded the average height of the men who registered there that day.
>
> Which would you expect to be true?
>
> The average height at post office A was closer to the national average than was the average height at post office B. *[Reversed]*
>
> The average height at post office B was closer to the national average than was the average height at post office A. *[Correct]*
>
> There is no reason to think that the average height was closer to the national average at one post office than the other. *[Equal]* (Well et al., 1990, Experiment 1)

Note that this problem is asking about how close the average of the sample is expected to be to the population. Well and colleagues term this an *accuracy* format, and 73% of participating psychology majors who had not taken statistics (*n* = 114) chose the correct answer, with only 20% choosing the Equal response. The authors note, however, that there are two simplifications working together as compared to the original hospital problem (Kahneman & Tversky, 1972): Not only does the task ask about a single sample rather than the distribution of many samples, but the task also draws attention to the central tendency rather than the tails. The above problem asks about *closeness* to the national average, whereas the hospital problem asks about how many days have more than 60% boys, in essence drawing attention to the *tails* of the sampling distribution. In their Experiment 2a, they created a new form of the problem which they term the *center* version that was still in terms of sampling distributions. The new text read as follows:

> Every day for one year, 25 men registered at post office A and 100 men registered at post office B. At the end of the day, a clerk at each post office computed and recorded the average height of the men who had registered there that day.
>
> Which would you expect to be true? (circle one)
>
> 1. The number of days on which the average height was between 5 feet 6 inches and 6 feet was greater for post office A than for post office B. *[Reversed]*
> 2. The number of days on which the average height was between 5 feet 6 inches and 6 feet was greater for post office B than for post office A. *[Correct]*
> 3. There is no reason to expect that the number of days on which the average height was 6 feet or more was greater for one post office than for the other. *[Equal]* (Well et al., 1990)

Well and colleagues (1990) also included a *tail* version for some subjects, in which the words "between 5 feet 6 inches and 6 feet" was replaced by "6 feet or more" (and thus #1 is *Correct* and #2 is *Reversed*). The authors found that the non-psychology-major participants in the *center* condition (*n* = 39, 50% Correct, 41% Equal) performed comparably to those in the *accuracy* condition (*n* = 75, 56% Correct, 40% Equal), and both center and accuracy versions outperformed those in the *tail* condition (*n* = 36, 8% Correct, 67% Equal). This shows that lack of understanding of the sampling distribution was not the only factor, since participants did nearly as well in the *center* condition.

However, further evidence of the role of sampling distributions emerged when Well et al. (1990) probed how participants had understood the task, after noting informally many misunderstandings of the problem when viewing written comments and talking with participants. Their Experiment 4 focused only on the 21 participants, drawn from an undergraduate psychology pool, who failed the *tail* sampling distribution version of the post office problem. Revealingly, 18 of these failing participants misstated the problem, with 11 of the 21 restating the problem as being about the *proportion of men in a single sample* rather than *proportion of means of many samples*. The authors further probed these participants' understanding by using a computer simulation to demonstrate the process of taking a sample of size 10 and plotting the mean on a separate graph. About half of the participants correctly predicted that the empirical sampling distribution with sample size 10 would have a lower proportion in the tail (greater than 71 inches) than in the population. All participants then went through the process of building up the sampling distribution for samples of size 10, and the interviewer made sure that participants understood the difference between the population and the sampling distribution and that they saw how the variability at size 10 was smaller than the population. However, even after this extensive demonstration of the reduction in variability of the sampling distribution of the mean for samples of size 10 in comparison to the population distribution, many participants (16 out of 21) still thought that a sampling distribution of the mean for samples of size 100 would have the same amount of variability as a sampling distribution for samples of size 10.

This study demonstrated two ways in which lack of understanding of sampling distributions can enter into judgements related to sample size. Participants may have so little concept of a sampling distribution that they are not even able to adequately represent the setup during encoding. Among those who have a basic understanding of the sampling distribution, however, some may not understand the role of sample size in sampling variability.

Sedlmeier (1998) focused on the distinction between frequency distributions and sampling distributions in his investigation of sample size tasks. For Experiments 1 and 2, he created *tail* single-sample frequency distribution versions of three classic tasks (the hospital problem, the word length problem, and the height problem) in Kahneman and Tversky (1972)— e.g., "*On a single day*, which hospital do you think was more likely to have more than 60% boys?"—and compared performance on those tasks with their original sampling distribution phrasing. Somewhat unusually, Sedlmeier used a within-subjects design and presented all of the tasks to his participants and focused his analysis on the *average* percentage correct of the hospital problem and the word-length problem (Experiment 1) or all three problems (Experiment 2).

This literature review has focused on the hospital problem due to its relatively clean presentation and its popularity in the literature. The word length problem may lead to different reasoning processes due to its convoluted context of comparing the average word length on lines vs. pages of a paperback book, a context that seems to make the sample space and the random process rather unclear (cf. Nisbett et al., 1983). Similarly, the height problem concerned the sampling distribution of medians, which have similar properties to the sampling distribution of means but may elicit different conceptions and are not the focus of this literature review. Given these substantial differences in surface features, it seems questionable to combine these three tasks into a single measure of participants' understanding of sample size, and Sedlmeier did not provide any evidence of consistency or reliability of the three items. Therefore, this discussion will focus on the hospital problem only, although the results for the other two tasks were qualitatively similar.

Sedlmeier (1998) first compared sampling distribution (*n* = 20) and frequency (*n* = 26) tail versions of the hospital problem among paid undergraduate students, and found substantially better performance on the frequency version (46% Correct, 42% Equal) compared with the sampling distribution version (20% Correct, 60% Equal). To address the concern that participants may simply not be understanding the problem setups, Sedlmeier followed Well and colleagues' (1990) example and provided simulation-based trainings to model both frequency and sampling distributions to the participants assigned to those groups, but without actually teaching the relationship between sample size and sampling variability. Trained performance on both types was improved compared to Experiment 1, but the frequency distribution version (US sample: 73% Correct, 27% Equal, *n* = 11; German sample: 60% Correct, 25% Equal, *n* = 20) still outperformed the sampling distribution version (US sample: 55% Correct, 36% Equal, *n* = 11; German sample: 30%

Correct, 50% Equal, *n* = 20). Although the number of participants in Experiment 2 was relatively small, the findings are consistent with the finding of Well and colleagues (1990) that training in the problem setup improves performance, and furthermore that sampling distribution problems remain more challenging than frequency distribution even after training. However, it should be noted that Sedlmeier (1998) only tested *tail* versions of items, which Well and colleagues (1990) found were more challenging for participants. As noted above, Well and colleagues found no substantial difference in performance between frequency distribution and sampling distribution items when both were phrased in *center* format (Experiment 2).

Sedlmeier (1998) argued that it was important to only test tail items because 9 out of 40 of Well and colleagues' (1990) participants in Experiment 3 who responded correctly to a center format item chose the reasoning the larger sample had "more opportunity for extremes and so the average height will be more variable" (p. 300). The larger sample does have more opportunity for extremes, but in fact the average height will be less variable and so this is incorrect reasoning. Sedlmeier (1998) argued that this is evidence that "asking about the middle of distributions might elicit response strategies that erroneously lead to correct answers" (p. 285). However, it should be noted that Well and colleagues' (1990) Experiment 3 did not have any possible response that corresponds to the idea of *balancing*, even though they found evidence of students reasoning that "larger samples provide more opportunity for large and small scores to balance out" (p. 309) in their Experiment 4. The "more opportunity for extremes" option was the closest response option given in

Experiment 3 that matches this conception of balancing—so it may be that some students did not see their correct reasoning among the response options and chose this as the closest answer. Moreover, it is unclear precisely what response strategy would erroneously lead to a correct answer. Balancing is discussed in more detail below in the section *Extreme values: more extreme values, swamping, and balancing* (p. 36).

In a wide-ranging literature review, Sedlmeier and Gigerenzer (1997) argued that the sampling versus frequency distribution distinction generally accounts for the differences in solution rates found across a wide variety of studies of sample size. As has already been shown, however, there are a variety of other task features that seem to greatly influence solution rates. Perhaps the most important is the difference between center and tail tasks (Well et al., 1990), which are often confounded with whether the tasks are in frequency or sampling distribution format. Lem and colleagues (2011) highlighted many other differences in tasks, such as extremeness of the tail cutoff (Bar-Hillel, 1982) and the sample size differential (Murray, Iding, Farris, & Revlin, 1987). It seems wise to moderate Sedlmeier and Gigerenzer's (1997) strong conclusion and instead to point out that there is indeed reasonable evidence that sampling distribution tasks are more difficult than frequency distribution tasks, and that this is one of the many factors that contributes to the differences found among different studies. Understanding the basic structure of sampling distributions is a necessary but not sufficient condition for understanding the nature of sampling variability. This result is unsurprising since, as Sedlmeier and Gigerenzer (1997) as well as statistics educators (delMas et al., 1999) have pointed out, neither evolutionary experience, everyday life, nor most formal schooling has prepared people to appreciate the properties of sampling distributions.

2.1.3 The size-confidence intuition

Students do seem to have some correct intuitions about sample size as well. In Well and colleagues' (1990) analysis of student reasoning of sample size tasks (Experiment 3), the majority of students who chose the correct answer to the sample size task (18 out of 40) chose the reasoning: "Larger samples are more likely to be similar to the population they come from than smaller samples, so if more men register on a day, the average height recorded should be closer to the average of the population" (p. 300). Indeed, Kahneman and Tverksy (1982) themselves raised the idea of using Socratic questioning to build on students' thinking about the "confidence in the results of a large sample" toward a correct solution on the original hospital problem (p. 500).

Where might such an intuition come from? Drawing on Piaget and Inhelder's (1951) studies of children's experience with randomizing devices, Nisbett and colleagues (1983) propose that children develop helpful *statistical heuristics* through their analysis of physical causation when applied to uncertain devices such as a spinner. This includes an intuition that larger random samples are more likely to be similar to the population they come from—the *size-confidence intuition* (Sedlmeier, 1999). Nisbett and colleagues (1983) argue that the size-confidence intuition plays a very general role in people's recognition that when generalizing from instances, more evidence is better than less. Similarly, Sedlmeier and Gigerenzer (1997) argue that such an intuition comes from evolutionary needs to weigh evidence, but that it only applies to frequency distributions since only those were accessible to humans during their evolutionary development.

If frequency distribution tasks could be used to prime participants to perform better on sampling distributions, this could be evidence of a facilitative role of the size-confidence intuition. In Experiment 3 of Sedlmeier (1998), *primed* participants (n = 16) were given an accuracy frequency distribution task involving emptying coins from urns with 10 or 100 coins in them, whereas the *unprimed* group (n = 15) received no such task. This prime was intended to elicit the size-confidence intuition. The primed group answered the prime question with 75% accuracy. Then, both groups were given a situation involving repeatedly emptying urns with 10 coins and 40 coins and were asked which urn would have more occasions of the number of heads being greater than 60%. Primed participants (38% Correct, 25% Reversed, 47% Equal) performed slightly better than unprimed participants (27% Correct, 0% Reversed, 73% Equal), but this difference was not reliable (p > .5). However, it is questionable whether this is a fair test of priming: participants were primed on a center frequency distribution but tested on a tail sampling distribution, which was found to be among the hardest by Well and colleagues (1990). Switching from center to tail may have caused confusion: All four of the primed participants who chose the larger sample (Reversed) on the sampling distribution problem had also chosen the larger sample (Correct) on the frequency distribution prime, whereas none of the unprimed participants chose the larger sample on the sampling distribution problem.

A similarly small and statistically unreliable priming effect was found by Brown (2015) in his study of students in a randomization-based introductory statistics course (Garfield et al., 2012; Zieffler & Catalysts for Change, 2017). Students (*n* = 74), who had not received instruction in sample size but had extensive exposure to sampling distributions, answered three accuracy frequency distribution questions either before (*Scaffolded*) or after (*Unscaffolded*) answering three graphical empirical sampling distribution items. Scaffolded students had, on average, 9 percentage points higher percentage correct on sampling distribution items than Unscaffolded students (*p*s > .3), but the reverse direction was even stronger—Unscaffolded students performed about 20 percentage points better on two of the frequency distribution questions than the Scaffolded group (*p*s = 0.05, 0.09). The idea that sampling distribution items could prime responses on the *easiest* type of frequency distribution item seems at odds with the size-confidence intuition and may implicate other learning mechanisms. Interestingly, in the Scaffolded condition the items were descriptively more correlated and had a more coherent mapping in adjusted multiple correspondence analysis than those in the Unscaffolded condition, which may be more suggestive of a role for the size-confidence intuition, but the difference in correlation matrices was also not statistically significant (*p* > .3). (Descriptive trends noted in this study should be taken with caution due to the low sample size and many comparisons made.) Additionally, all three frequency distribution items focused on the center whereas three sampling distribution tasks focused on differing regions, exhibiting similar problems to Sedlmeier (1998).

Lem (2015) addressed this limitation in her study of Dutch M.Ed. students' (*n* = 65) intuitions about the hospital problem. In her Experiment 1, she administered the hospital problem consistently in tail format, but in both frequency and sampling distribution format, with participants randomly assigned to frequency-first and sampling-first problems. Participants received the second problem immediately after the first problem. She found essentially no differences by either order (frequency-first vs. sampling-first) or for type of distribution (frequency or sampling). Some of this lack of effect may be due to the high level of education of the participants and the fact that participants had received the same introductory statistics course. Statistics training has been shown to boost people's comprehension of sample size problems (Fong et al., 1986). Furthermore, the problems are so similar that administering them right afterwards may have led people to respond the same way to both problems by analogy.

To further probe the potential role of the size-confidence intuition in frequency and sampling distribution problems, Lem (2015) employed a visual distractor task to see whether loading working memory would force intuitive processing and suppress inhibition when answering the hospital problem in a study of undergraduate Dutch psychology students (*n* = 181) who had taken an introductory statistics course. She did not find any reliable evidence of any effect of the dot pattern, although descriptively the accuracy was lower for both types of task when there was the distractor task. Surprisingly, there was poor accuracy even in the no-load condition for the frequency task (24% correct) when compared with the sampling task (23.1% correct). Lem (2015) surmised that some of the difference may be due to the translation into Dutch, but it is not clear why the performance on the frequency task would be so different than Seldmeier (46% correct on frequency; 20% correct on sampling).

Thus, the most direct evidence of something like a size-confidence intuition remains the chosen reasoning about large samples being more representative given in Experiment 3 of Well and colleagues (1990). The authors caution that this reasoning may be a “fuzzy heuristic” that is not always consistently applied, and indeed some of their participants were as vague as to say simply something like “bigger is better.” While not incorrect, this type of reasoning may be very fragile and not stably evident across a variety of tasks and circumstances. Lem (2015) theorizes that the size-confidence intuition may not be a very strong intuition and may come through education and experience, consistent with the accounts in Piaget and Inhelder (1951) and Nisbett and colleagues (1983).

2.1.4 Sample-Population Ratio

Unfortunately, the idea of larger random samples being more representative does not distinguish precisely *why* they will be more representative. Are larger random samples more representative because of the Empirical Law of Large Numbers, or because they cover a larger proportion of the population? Once the population is sufficiently large, additional population size negligibly influences inferences. However, there is some evidence that people may be attending more closely to the sample-population ratio than to the sample size alone, which may be another undesired generalization of proportional reasoning.

Bar-Hillel (1979) explored undergraduate participants' thinking about sample size and sample-population ratio in a series of experiments. First, she expanded the polling problem to include population sizes in her Problem 6 (Firm A: 50 out of 50,000 vs Firm B: 100 out of 100,000). Although 50% of the undergraduate participants (*n* = 24) correctly chose the larger absolute sample, 29% of participants indicated that they would have equal confidence and provided justifications such as "proportionately the same number are being sampled" (p. 252). This response may have been cued due to problem phrasing that made the proportionality rather salient ("Both firms are sampling one out of every 1000 voters", p. 251).

However, a different sample of undergraduates saw her problem 7, which had the same population sizes and with both firms sampling 1000 participants without calling attention to the now-different proportionality. Bar-Hillel coded "same" as the normatively correct answer, but technically the finite population correction would still very slightly favor the situation with the smaller population size. Therefore, the fact that 62% of participants (*n* = 21) preferred the firm with the smaller population is not incorrect. However, she did note that these participants often justified their reasoning with proportional reasoning such as "because B is interviewing a bigger percentage of people than A", and evidence of sample-population ratio reasoning is stronger here than in problem 6 since the proportionality was not highlighted for participants. Note that these polling problems were asking about subjective confidence, which may lead to different types of judgments than asking about likelihood (Brown, 2015; Hullman et al., 2015).

Some more evidence for Bar-Hillel's (1979) hypothesis that sample-population ratio is driving participants' preferences was provided when she gave undergraduate participants (*n* = 57) nine different pairings of sample size and population size and asked participants to rate their "accuracy" on a scale of 1 to 10. The sample sizes and population sizes varied in such a way that the ranking by sample size and sample-population ratio were distinct (*r* = -.11) so that participants' preference for absolute sample size or ratio could be distinguished. The situation was such that sample sizes were all so large that normatively the finite population correction would be so small that population size should be essentially irrelevant. Participants' median ranking correlated much more highly with the sample-population ratio than with the sample size, and 39 of the 57 participants had more high correlations in the same direction. For example, participants gave a median accuracy rank (higher rank = greater accuracy, max rank = 9) of 4 to a sample of 100,000 out of 100,000,000 (ranked 8.5 in sample size and 2.5 in sample-population) while giving a median rank of 7 to a sample of 50 out of 100,000 (ranked 1.5 in sample size and 8.5 in sample-population). These results do provide further evidence that participants were attending to sample-population ratios. However, within-subjects manipulation of this factor may have led participants to attend more to this factor than they might otherwise, as several researchers have pointed out (Kahneman & Tversky, 1982; Pollard & Evans, 1983).

Although the contribution of sample-population ratio to students' understandings about sample size has not been widely studied, the strong findings from Bar-Hillel (1979) indicate it may require attention. Bar-Hillel posits that attention to this ratio may come from people's many everyday experiences sampling from very small populations where the ratio is more of a normative concern, such as dishes at a restaurant.

Well and colleagues (1990) also found an interesting type of reasoning where participants reasoned in a correct way that depended on sampling without replacement. In their Experiment 4, some participants (those who had failed the post office problem on a screener, $n = 21$), after viewing a computer demonstration of how the empirical sampling distribution is generated, commented that a sampling distribution of means of 10 adult male heights would have less variability than a population distribution of heights because there would *not be enough of the most extreme scores* to fill the entire sample, and therefore the means would not be as extreme as the most extreme scores of the population. Their task, which had a population of only 400 adult male heights, appears to have been presented in a way where participants may have thought sampling was without replacement. Additionally, three of these participants, after viewing the lower variability of sample means for samples of size 100, explained that the larger sample is more representative because it includes a larger proportion.

Viewing the sample primarily in terms of the proportion of the population may relate to what Saldanha and Thompson (2002) refer to as an *additive* conception of sample as simply a subset of a population. These students may not be confusing frequency and sampling distributions as seriously as the Well and colleagues' (1990) participants who could not even restate a sampling distribution problem correctly. However, they do not possess a rich *multiplicative* image of sampling as a repeated, "quasi-proportional" process whereby the repeated samples vary in ways related to their size. According to Saldanha and Thompson, those who view the sample simply as a subset may focus on the relative size of that subset, rather than on the properties of repeated sampling.

2.1.5 Extreme values: more extreme values, swamping, and balancing

One reason for the difficulty of tail problems is that participants may have difficulty reasoning about extreme values. For instance, in Experiment 3 of Well and colleagues (1990), nearly all of the participants choosing the *Reversed* options ($n = 13$) chose the explanation "If a larger number of men register on a day, there will be more opportunities to have extremely tall or short men and so the average height will be more variable." Chance and colleagues (2004) note informally that a "common misconception" about sampling distributions is that "sampling distributions for large samples have more variability" (p. 302), perhaps following a similar line of reasoning.

However, participants can also reason correctly about extreme values. Crucial to this type of reasoning is understanding how extreme values in the sample affect the mean at different sample sizes. Following the terminology in Well and colleagues (1990), *swamping* is the phenomenon that at larger sample sizes, any individual value of the sample contributes much less to the mean and therefore extreme values will not affect the mean as much. In their Experiment 3, Well and colleagues (1990) found that 12 of the 40 participants who chose the correct answer chose the explanation, "If a larger number of men register on a day, the average height recorded is less likely to be influenced by a few extremely tall or extremely short men" (p. 300). After receiving training in Well and colleagues (1990) Experiment 4, several participants spontaneously volunteered a swamping explanation. Chance and colleagues (2004) also observed this explanation, with one participant explaining, “if you keep taking averages, the outliers are going to actually be less, um, have less big effect on your data. So you’re actually always dropping out those outliers [...] so it will get more and more narrower” (p. 310).

Another correct type of reasoning based on extreme values is *balancing*: Larger samples are more likely to have extreme values on both sides of the mean, and therefore these will balance out. After the training in Well and colleagues’ (1990) Experiment 4, some unspecified number of participants displayed this type of reasoning. Balancing reasoning may have also have occurred in Experiment 3, but none of the options corresponded with an idea of balancing. No other literature discussing this type of reasoning about sample size was found, despite the emphasis in the mathematics and statistics education literature on students’ understanding of the mean on the mean as a center of balance (e.g., Franklin et al., 2005; Mokros & Russell, 1995; O’Dell, 2012).

2.1.6 Intuitive Homogeneity

The above research has focused on the role of sample size, but sample size and variability interact in determining the variability of the sampling distribution. Many of the classic tasks include some kind of cue to variability to make sure participants do not view the process as deterministic, since without variability all the problems would require is simple proportional reasoning and the “equal” choice would be trivially correct (cf. Fischbein & Schnarch, 1997). For instance, in the classic hospital problem, participants are reminded that "The exact percentage of baby boys, however, varies from day to day. Sometimes it may be higher than 50%, sometimes lower" (Kahneman & Tversky, 1972). In Well and colleagues' (1990) weighing problem, they discussed features of the scale: "Although the scale is not completely accurate, it is equally likely to read high as it is to read low. However, it is never off from the actual weight by more than 25 grams" (p. 293).

Indeed, Nisbett and colleagues (1983) provide evidence that people make judgements based on intuitions about the homogeneity of groups. Their participants were told that they were explorers on a fictitious island, and encountered a number of islanders (n = 1, 3, or 20, manipulated between subjects), all of which were obese, and a number of specimens of an element, "floridium", all of which were conductive (n = 1, 3, or 20, manipulated between subjects). Participants were asked what percentage of the populations they expected to have that same property—i.e., what percentage of islanders are obese (high context-implied variability), and what percentage of pieces of floridium are conductive (low context-implied variability)? On average, subjects estimated high population percentages for all three sample sizes for conductive floridium, but increasing percentages by sample size for obese islanders (M = 35% for n = 1 participants, M = 55% for n = 3 participants, M = 75% for n = 20 participants).

Obrecht, Chapman, and Gelman (2007) analyzed intuitive comparisons of product ratings for two products among college students. Their stimuli displayed the means, standard deviations, and number of ratings for each product; the standard deviations and number of ratings were the *same* for any given pair of products whereas there was always a *difference* between means. Participants then were asked how confident they were that the product with the higher mean rating was in fact better than the product with the lower mean rating. The students adjusted their confidence ratings in relationship to all three factors. However, students responded much more strongly to the mean differences than to the sample size, and barely responded to differences in standard deviation, relative to the degree of power difference. People with statistical training did respond somewhat more to sample size, but still very little to standard deviations.

However, the picture gets more complicated in a follow-up study where context was explored in more depth using the same paradigm (Obrecht, Chapman, & Suárez, 2010). First, they replicated the findings of Nisbett et al. (1983) in the context of two groups, again drawing on the conductivity and body weight on an island context but now with means and sample sizes given. Subjects were sensitive to sample size, context-implied variability, and the mean difference. The researchers then examined the interaction of context-implied variance based on the body weight and conductivity with the actual *SD* of the sample. They found that participants were sensitive to variance when *both* the context and the within-group variance *matched*—but were not sensitive to within-group variance when the two were *mismatched*, e.g. if floridium showed high variability in conductivity. Interestingly, subjects were more sensitive to sample size in this experiment than they were in Obrecht et al. (2007). The ratio of the sample sizes of the two groups was larger in the 2010 study than in the 2007 study, and indeed the extremeness of the difference between sample sizes has been found to make sample size more salient and to increase the proportion of correct responses (Murray et al., 1987).

2.1.7 Summary of Sample Size Research

Previous reviews of sample size tasks have focused on task and participant features that lead to differences in percentage correct found in the literature (Lem et al., 2011; Pollard & Evans, 1983; Sedlmeier & Gigerenzer, 1997). In contrast, this review has focused on the reasoning and conceptions revealed by different sample size tasks and by asking participants for their reasoning. Eight major conceptions about sample size and sampling variability were identified (Table 2.1).

Table 2.1
Conceptions about sample size and sampling variability

Response Type	Reasoning	Description	References
Correct	Swamping	Average influenced less by extreme values	Well et al., 1990
Correct	Size-confidence intuition	Large samples more like the population, so more likely to have mean near population	Well et al., 1990
Correct	Balancing	Larger samples provide more opportunity for large and small scores to balance out	Well et al., 1990
Correct	Sample-population ratio	Larger samples capture a larger proportion of the population	Bar-Hillel, 1979
Equal	Population similarity/proportional reasoning	Sample size is not a factor because the problem deals with averages	Kahneman & Tversky, 1972; Well et al 1990
Equal	Frequency confusion	Proportion of scores expected in range of *frequency* distribution same regardless of sample size	Well et al 1990
Reversed	More opportunity for extremes	More opportunity for extremes with larger sample size and therefore the mean is more variable	Well et al 1990
(Context-dependent)	Intuitive homogeneity	Sample size is relevant when the context implies substantial variability in the population	Nisbett et al., 1998; Obrecht et al., 2010

The large literature on sample size reasoning reveals the inadequacy of existing theories of sample size reasoning to account for the diverse ways of thinking about sample size, population variability, and sampling variability. Clearly there are many other observed participant conceptions besides population similarity (i.e. the representativeness heuristic), which Kahneman and Tversky (1972) posited as a general principle of human judgment under uncertainty. Difficulties with sampling distributions, posited by Sedlmeier and Gigerenzer (1997) to be the primary driver of sample size neglect, also appears to be only one of several possible task features that lead to correct reasoning (e.g., Lem et al., 2011).

Instead of seeking a single, simple theory for participants' understanding of sample size, it may be more useful for future theories to examine how each of the eight conceptions listed in Table 2.1 arise. What kinds of tasks elicit each of the eight sample size conceptions? How do various types of statistical training and experience change which types of sample size conceptions participants tend to have? What are the roles of individual differences in statistics education, math ability, or cognitive reflection (Frederick, 2005)?

From a pedagogical perspective, what is the best pathway from an incorrect conception to a correct conception? Given the persistence of the misconceptions found in the literature, changing from an incorrect to a correct conception—or perhaps to multiple correct conceptions—is likely to be a daunting task. The next section reviews recent literature on conceptual change and its relevance to sample size interventions.

2.2 Conceptual Change for Sampling Variability

There are numerous factors that contribute to changing students' understanding of any statistical concept. Yet there are persistent inconsistencies in students' reasoning about sampling variability and the Empirical Law of Large Numbers even after intensive intervention informed by the conceptual change literature (delMas et al., 2006). Why is sampling variability difficult to learn? A look at more recent conceptual change literature suggests that students' difficulties may be due to lacking an understanding of the *mechanism* of the Empirical Law of Large Numbers—and this mechanism is particularly difficult to learn because it is an *emergent* property of means that is not readily apparent from experience with samples. This section first reviews conceptual change literature related to causality, mechanism, and emergence, and concludes with an examination of previous studies of conceptual change of sampling variability through this lens.

2.2.1 Causality, Mechanism, and Emergence

Why are *causality*, *mechanism*, and *emergence* important? These are not mentioned in classical approaches to conceptual change, which emphasize creating cognitive conflict in students in order to lead them to see the inadequacies of their existing concepts and to therefore replace them with new, correct concepts (Posner, Strike, Hewson, & Gertzog, 1982). However, research in reading comprehension suggests that old knowledge, even when the learner understands that the information is no longer correct, remains in long term memory and always has the potential to be reactivated (reviewed in Kendeou & O’Brien, 2014). This classical approach also does not account for the fragmentation in understanding science education researchers have observed (Vosniadou & Skopeliti, 2014) when students insert a learned fact (“the earth is round”) into a non-normative framework (“the earth is flat”), resulting in a chimerical understanding (e.g. “the earth is round like a pancake”). Thus, converging lines of evidence have indicated that knowledge can not simply be replaced.

The Knowledge Revision Components (KReC) framework (Kendeou & O’Brien, 2014), drawn from the reading comprehension literature, suggests that the best-case scenario for conceptual change is when new correct information is *learned and activated* *at the same time* as old incorrect information, and is *integrated* with the old information. One mechanism for integrating the new and old information is a refutation text, which states the old information, explicitly refutes it, and then states the new information. Then, whenever the old or new information are activated by memory, the memory representations associated with both pieces of information undergo *competing activation.* If the new information can be given a rich and memorable explanation, the new information may be able to win out in memory so that the old incorrect information does not interfere with comprehension (Kendeou & O'Brien, 2014). KReC proposes the necessity of building a rich network of cognitive connections around the new information. Causal explanations are one method of building such a network (Kendeou et al., 2013). Even if a person *understands* and is *convinced* of the new conception in ways consistent with the approach given by Posner and colleagues (1982), this is not sufficient for preventing old conceptions from interfering later on. The old, incorrect knowledge still has the potential to be activated indefinitely, and the new information must be able to compete successfully with it for the conceptual change to stick (Kendeou & O'Brien, 2014).

The causal explanations studied in Kendeou et al. (2013), however, are short phrases supporting a refutation text, which seems inadequate to convey the complex, inter-level, and context-sensitive phenomenon of the Empirical Law of Large Numbers. What might an ideal causal explanation consist of? Literature in the science education about similarly complex phenomena has targeted *mechanistic reasoning.* In their synthesis of studies of scientific mechanistic reasoning, Russ et al. (2008) describe mechanistic reasoning as nonteleological, causal, built from experience, and descriptive of underlying structure. In the terms of Russ et al. (2008), causal understanding may just consist of knowing that when A happens, B happens; but mechanistic reasoning additionally includes understanding *how* A causes B. Given how different contexts and structural features may cue quite different reasoning in isomorphic situations regarding sampling variability (e.g., Wagner, 2006), an understanding of underlying structure may allow normative understandings of the Empirical Law of Large Numbers to compete more successfully with misconceptions, because it successfully integrates the Empirical Law of Large Numbers with the variety of structural features that people can observe in problems. Russ et al. (2008) draw on a definition of *mechanism* by three philosophers of science: "Mechanisms are entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions" (Machamer, Darden, & Craver, 2000, p. 3). Thus, students' mechanistic reasoning about sampling variability can be explored in terms of the entities, activities, organization of entities and activities, and regular changes that they articulate while interacting with sampling variability.

Notably, the mechanism of sampling variability does not operate in a sequential and linear fashion, and this may make it particularly difficult to understand (Chi et al., 2012). Chi and collaborators (2012) distinguish between *sequential* processes such as blood flow through the circulatory system and *emergent* processes such as diffusion. Sequential processes have differentiated roles (the heart pumps blood) wherein different parts function together to achieve the overall function of the system through a series of additive events (oxygenating blood in the lungs, pumping it out to the arteries, retrieving through the veins, and repeating). Sequential patterns are generally conducive to narrative explanation which comes naturally to most students. In contrast, emergent patterns tend to have uniform parts (individual molecules of water and dye) who operate similarly to each other (moving randomly and colliding) in a way that collectively sums to produce the overall aggregate pattern (the color of the dye spreads out in a glass of water). Participants who received general training in emergent processes benefited more from a computer simulation of diffusion than those who did not (Chi et al., 2012).

Pratt and colleagues (2008) have analyzed how students have trouble with emergence in the context of connecting *local* uncertainty with *global* patterns of invariance in randomness. The convergence of the sample mean to the population mean under the Empirical Law of Large Numbers is an emergent process. Each case that is randomly sampled is weighted equally in the sample mean (*equal status*), and individual cases can be far away from the mean (*disjoint*), and it may even be impossible for a single value to equal the population mean (e.g. a Bernoulli distribution; *local goals*); however, the *net effect* of the *entire collection* of all values is that additional cases tend to bring the sample mean closer to the population mean. Thus, lack of understanding of emergence more generally may lead to profound misconceptions on the part of students, and this may in part be responsible for students' frequently slipping between frequency and sampling distribution interpretations (Chance et al., 2004; Well et al., 1990).

2.2.2 Studies of Conceptual Change in Sampling Variability

The previous section argues that successful conceptual change in student understanding may involve supporting students' ability to construct a mechanistic causal explanation as well as supporting students in understanding emergence. How have prior studies of conceptual change in sampling variability supported mechanistic reasoning and understanding of emergence?

In perhaps the first systematic attempt at examining interventions with students' application of the Empirical Law of Large Numbers to samples, Fong et al. (1986) compared different levels of training in the Empirical Law of Large Numbers: the *control* group received no instruction; the *demand* group was simply informed that there was a useful principle called the Law of Large Numbers; the *example training* group had a detailed introduction to the Empirical Law of Large Numbers provided through four worked applied examples; the *rule training* group received a four-page description of background statistical concepts, a general statement that frequency distributions increasingly resemble population distributions as the sample size increases, a worked example involving sampling blue (70%) and red (30%) gumballs from an urn, and a live demonstration where an instructor sampled from the gumball machine at sample sizes 1, 4, and 25 and wrote down the proportion blue; and the *full training* group received a combination of both the example and rule trainings. Fong et al. found that the full training group performed the best on posttest items, and that the example training and rule training groups performed similarly to each other and much better than the control and demand groups.

Fong et al. (1986) did indeed convincingly demonstrate that, at least in the short term, that understanding of sampling variability is sensitive to instruction. However, the between-subjects design of Fong et al. (1986), with only a posttest, make it unclear exactly what or how conceptual change may be occurring in these individuals. The authors simply provided evidence that participants with certain trainings can perform better than others with no training immediately after being trained. Additionally, Fong et al. (1986) only assessed participants' ability to apply the Empirical Law of Large Numbers to certain textual problem structures, and it is unclear if they would be able to apply it to statistical contexts such as sampling distributions and power. Finally, as Sedlmeier (1999) has pointed out, it is normatively not clear whether the Empirical Law of Large Numbers should be applied in many of the contexts Fong et al. present to their subjects, which by design do not involve any reference to random sampling.

Although no conditions in Fong et al. (1986) seem to provide any explicit support for mechanistic reasoning, the rules and full trainings do provide a live demonstration of the Empirical Law of Large Numbers and the process of repeatedly taking samples. This may provide some support for connecting frequency and sampling distributions and provides at least a minimal opportunity for observing the process of emergence, though the contribution of these factors to the authors' findings is not possible from the data they provide.

In a similar follow-up study, Fong and Nisbett (1991) assigned participants to either an *ability testing training* group, who were trained with a general one-page introduction to the Empirical Law of Large Numbers and then examples in ability testing, and a *sports training group*, who were similarly trained but with examples in sports performance. Not much information is provided about the one-page introduction, so it is unknown whether it provided any causal explanation. However, the inclusion of a causal explanation seems unlikely given what was provided in Fong et al. (1986), the lack of any mention in the article, and fact that the training seemed greatly condensed and without the demonstration component of Fong et al. (1986). Both groups demonstrated sustained high performance after two weeks when tested on items in the same context as training (e.g., when ability testing training participants received ability testing items), but there was a noticeable decrease in performance after two weeks when tested on items in a different context (e.g., when ability testing training participants received sports items). The decay of transfer outside of the trained context provides some more evidence that the understanding imparted by the training—now consisting entirely of direct instruction in a rule and practice applying it—may not be supporting very deep or comprehensive conceptual changes in understanding of sampling variability. Furthermore, as Sedlmeier (1999) has pointed out, performance was assessed using a "mean statistical reasoning score" that is difficult to interpret and without any psychometric properties assessed. Moreover, the details of the posttest items in both Fong et al. (1986) and Fong and Nisbett (1991) are relatively scant, so it is not clear what domain of sample size tasks participants are performing well on.

Although, as noted above, Sedlmeier (1999) fails to provide clear evidence for a central role of the size-confidence intuition in correctly solving sample size tasks, his intervention was relatively successful in training participants to solve sample size items similar to the hospital problem. Drawing on his theory that *adaptive algorithms* are more well-tuned to some formats than others, his proposed intervention method is to provide tools for people to translate from frequency to sampling distribution formats. Sedlmeier does this by creating a computer program that shows the process of sampling by using a depiction of an urn in which balls are mixed and then drawn with or without replacement. The frequency distribution is displayed and the connections of this to the sampling distribution are shown. He theorizes that this model is analogous enough to experience—perhaps because it is analogous to the randomization devices discussed by Piaget and Inhelder (1951)—that it will allow people to translate the sampling distribution into a frequency distribution and thereby to activate the size-confidence intuition.

Sedlmeier's (1999) intervention showed evidence of success in training people on sampling-distribution tasks that are isomorphic to the hospital problem or to *center* problems such as the weighing problem (Well et al., 1990). Even after one week and five weeks had elapsed, the German undergraduate participants' (*n* = 20) average percentage correct on forced-choice tasks and to sampling distribution construction tasks was around 80%, whereas performance had ranged from 20% to 50% on the pretest. Apparently Sedlmeier has created a well-designed and successful intervention for teaching how to solve three kinds of sample size tasks (isomorphs to the hospital problem, isomorphs to *center* versions of the post office problem, and understanding a few features of constructing subjective sampling distributions). However, it is unclear what features of the intervention were helpful, and no comparison groups were part of the design. Thus it is not clear whether the improvement was indeed due to participants mapping sampling distribution problems onto frequency distributions in a way that activated the size-confidence intuition. Besides the flexible urn model, other features of the intervention that could have led to the successful results are:

- A clear explanation of the different parts of the sampling process and the connections between them
- A rich technological simulation environment
- Extensive practice and visualization of very similar tasks (all the choice tasks were concerned with the tail region of *60% or more* or the center region of *between 40% and 60%*)
- Explicit instruction in the rules to follow in evaluating sample size tasks. For instance, participants saw the following instruction after one demonstration of the sampling distributions implied by the hospital problem: “Generally speaking: The larger the sample size the more likely the proportion calculated from that sample is close to the true proportion (and the less likely this proportion lies far away from the true proportion)” (p. 132)

Although it is difficult to directly compare the results of Fong et al. (1986) and Fong and Nisbett (1991) to Sedlmeier (1999) given the different tasks and the lack of detail in the Fong studies, these features in Sedlmeier (1999) do seem to provide more support for understanding the emergence of sampling variability. Although Sedlmeier's representations and intervention tasks do not make clear the mechanisms of sampling variability, they may provide more support for students understanding the relationship between frequency and empirical sampling distributions and observing the emergence of sampling variability upon repeated samplings. In contrast, the rules and full trainings in Fong et al. (1986) did not display sampling distributions or provide as much sequence, since participants only noted the proportions given from random samples from the gumball machine.

In a similar series of interventions by delMas and collaborators (Chance et al., 2004; delMas et al., 1999, 2006), students were provided even more opportunities to experience the causal relationships between different contributing factors to the shape and spread of sampling distributions using *Sampling Distributions* and its successor, *Sampling SIM* (delMas, 2002). These open-ended microworlds allowed students to control many input parameters of sampling distributions, including drawing their own population distribution or selecting from preset distribution shapes, selecting sample sizes, and from there drawing random samples to create empirical sampling distribution. Although again the representations did not provide direct support for mechanistic reasoning, the ability to manipulate the input factors may have provided opportunities for students to start developing causal reasoning. Additionally, similar to Sedlmeier (1999), students could see the frequency distributions and how they connected with the sampling distribution, which might help support their understanding of emergence. When combined with a classical conceptual change approach (Posner et al., 1982) that confronted students with their misconceptions after the pretest, this intervention appeared to be quite successful and encouraging most participants to choose histograms with smaller variability for larger sample sizes (delMas et al., 1999).

It is not clear, however, how much the success in training the histogram task in delMas et al. (1999) transfers to other contexts. In an unpublished follow-up study of students who had undergone the Sampling SIM (delMas, 2002) intervention, delMas and colleagues (2006) tested students on both the histogram tasks and some classic sample size tasks. The results were not encouraging for the ability to transfer from the histogram tasks to other contexts. All participants did quite well (87%) on the rock weighing problem (Well et al., 1990) that was phrased in accuracy form, but this is unsurprising because even uneducated participants can do fairly well on this problem already (Well et al., 1990) and it was phrased in terms of a repeated measurement context which may be more conducive to properly understanding the role of center and variation (Konold & Harradine, 2014). However, participants fared less well on other problems. Even students who had consistently chosen correct or reasonable histograms were below chance levels (29.2% correct) on the sampling distribution tail version of the post office problem (Well et al., 1990), well in line with the proportions found in the sample size neglect literature among naïve undergraduates.

One criticism of prior interventions for developing understanding of sampling distributions is that they do not appear to provide a causal explanation for the mechanism of the Empirical Law of Large Numbers. Ultimately, these prior interventions rely on demonstrations to show *that* sample size has such an effect and this weaker conceptual network may be less general than a mechanistic causal explanation, and may not readily transfer outside contexts that have very similar structure. Despite the successes of Sedlmeier's (1999) intervention on sample-size tasks, including solid retention results several weeks after the training, the training and testing tasks are closely isomorphic sample size problems based on the hospital problem. It is not clear whether participants would be able to use them successfully in more complex statistical contexts such as comparing groups. Similarly, it is notable that participants in delMas and colleagues' simulation-based training (delMas et al., 2006) were able to improve substantially in histogram comparison tasks analogous to the training but not to transfer this understanding to classic sample size problems. One participant in Chance and colleagues' (2004) work exemplifies the lack of mechanistic or even causal understanding:

> Because I just remember learning in class that it goes to a … when you draw, it goes to a normal distribution, which is the bell-shaped curve. So, I just look at graphs that are tending toward that … That's just how … I don't know like the why, the definition. That's just what I remember learning. (Chance et al., 2004, p. 309)

This student seems to have some correct declarative knowledge of the Central Limit Theorem (and thus of the smaller variability and increasing bell-shape of the sampling distribution of the mean at larger sample sizes), but was only able to justify their reasoning by an appeal to authority.

Many interventions do a good job of distinguishing and clarifying the levels of sampling and sampling distribution (Chance et al., 2004; Sedlmeier, 1999; Well et al., 1990), but do not actually highlight the *inter-level attributes* that provide the *mechanism* for creating to the pattern. Students may be able to learn macro patterns successfully while still having deep misconceptions about emergence, and so these interventions may not be revealing and addressing these misconceptions (Chi et al., 2012).

2.3 Supporting Mechanistic Understanding of Sampling Variability

This review now refocuses on the literature related to potential targets for supporting mechanistic understanding of sample size's effect on sampling variability: the size-confidence intuition, swamping, and balancing. The size-confidence intuition relates to literature on students' understanding of what researchers have called the Law of Large Numbers, but little in this literature seems to provide support for a mechanism. Both swamping and balancing emerge as mechanisms of the mean; while a great deal of research has been done on students' mechanistic understanding of the mean, this has been widely applied to sampling variability. Moreover, it is argued that swamping and heaping fail to capture important aspects of the mechanism of sampling variability since these principles do not depend on random sampling. Another mechanism, here termed *heaping*, captures more of the sampling variability phenomenon and is hinted at in research into students' understanding of sample spaces. For each of these potential targets, the implied mechanism is described, and then prior research is examined for how these targets could be supported.

2.3.1 The size-confidence intuition: true probability and experimental outcomes

One of the frequently cited reasons for participants to make correct choices in sample size tasks is the size-confidence intuition that larger samples will be more representative of the population (e.g., Sedlmeier, 1999; Well et al., 1990). What reasoning could support this intuition to support mechanistic reasoning on sample size tasks? Sedlmeier (1999) explicitly intended to support the intuition with ambiguous results reviewed above. However, the emergence of long-run patterns in sampling distributions is closely related to the emergence of long-run convergence of the proportion of an event occurring to the true probability under the frequentist definition of probability. There is a fair amount of research on elementary and middle school students' understanding of this convergence, which the authors call the Law of Large Numbers but which this thesis terms the *Empirical* Law of Large Numbers to distinguish the practical phenomenon in finite samples from the asymptotic theorem (see Chapter 1).

In their qualitative study of 10–11-year-old's reasoning about simulations of non-standard unknown dice in the *InferenceMaker* software, Pratt and colleagues (2008) hoped to draw students' attention to the relative stability of proportions at large sample sizes. The instructors could create a hidden die that could have any number of sides and any numbers on those sides—for instance a six-sided die with no ones or twos but three sixes. The children could "roll" the die as many times as they wanted within the software and view individual outcomes, a pie chart showing the relative frequency of each number on the die, or a pictogram representing how frequently each number occurred. The goal was to figure out how the die was configured based on repeatedly rolling it.

One efficient strategy would be to simply roll the die several thousand times and infer the proportions based on that; however, the participants in Pratt and colleagues (2008) did not follow this strategy. Students took their samples incrementally, examining the distribution of numbers, adding more outcomes to the sample, and then noticing changes as they added more and more outcomes. At a certain point students might start a whole new sample and compare the graph of the new sample with the old one. However, neither of these approaches appeared to help students understand that larger samples will be more reliable. After students drew conclusions based on a larger sample, the researchers would draw a new sample of 10 to compare and students were often thrown off by this new, smaller, and unreliable sample. Students attended more to changes even in larger samples (e.g. noticing that the fives were "growing") rather than the long-run stability, and appeared to be constantly frustrated at the apparent lack of stability. One student commented that as the sample got larger it was "just getting too stupid." Pratt and colleagues commented that this confusion may be because of the emergent nature of the stability. Since collecting a new sample of a small size would always show instability and a different pictogram or pie chart, it is only by focusing on the aggregate perspective for a large sample that the stability can be viewed. As the authors noted, some of this problem may be due to the limited annotations in *InferenceMaker*: students could save pie charts or pictograms but these did not have labels with the sample size on them.

A useful feature of the tasks in Pratt et al. (2008) was that the hidden die appeared to be successful in engaging students to thinking about deviations, and to connecting what they were observing in samples with the hidden structure of the dice. Part of the mechanism of sampling variability depends on the sample space and students' attention appeared to be successfully directed there in this study. Another strength is that students did appear to be able to view long-run stability from the relative lack of movement in the pie chart representations. However, their representations failed to make sample size salient, and thus it would be hard to accurately view the mechanisms of the Empirical Law of Large Numbers.

Studying a similar population of 10–12 year old students, Ireland and Watson (2009) attempted to bridge ideas of theoretical and experimental more explicitly through a structured class activity on coin tosses at different sample sizes and discussing the effect of sample size on the closeness of experimental outcomes to the theoretical probability. The first lesson involved physical coin tosses and the second lesson used TinkerPlots™ (Konold & Miller, 2017) to simulate large numbers of coin flips. This prior instruction appeared to support these students in more successfully noticing and incorporating the Empirical Law of Large Numbers when testing an unknown die to see whether it was "fair". Student quotes such as "the dice is never going to be the same but the more you do, the more the same they are going to be" (p. 353) suggest stronger understanding on the part of these students and an appropriate attention to the aggregate view.

This study also included explicit comparisons at different sample sizes, which presumably would be more supportive in making sample size more visible. However, four of the eight participating students had difficulty drawing these connections, and expected short-term fairness as well (cf. the "law of small numbers", Tversky & Kahneman, 1971). Moreover, Ireland and Watson raise the concern that the automatic rescaling that TinkerPlots™ does upon displaying larger trials seemed to be crucial to students seeing the leveling-off—or not, when the die was loaded—upon seeing larger sample sizes. It is not clear that the students understood this rescaling or would be able to reason correctly in its absence. Also, the researchers did not challenge students with new, small samples as Pratt and colleagues (Pratt et al., 2008) did, which might provide more information about the limits of their understanding of the Empirical Law of Large Numbers. Like many sample size and sampling distribution activities, neither the instruction nor the simulation highlights a mechanistic reason *why* large samples may lead to distributions being more likely to be similar to the population.

In their study of 11–12 year old children exploring possibly biased dice using *Probability Explorer* (Stohl-Drier, 2000), Lee and colleagues (2010) found more mixed evidence of successfully coordinating reasoning within and across multiple samples. One student who saw persistent, but low-magnitude, unevenness even after 906 trials continued to assert that the die was fair. Other groups were more successful in detecting that their die was biased, and some quotes do point toward an understanding of the Empirical Law of Large Numbers—based on a large trial, one student said, “these percentages help us in, like, determining what the probability is moving to”. However, large samples were not consistently used, indicating that the “size-confidence intuition” may not have been strongly present. Again, no reason *why* large trials might be more reliable was either offered to students or volunteered by them.

In contrast to many of these studies of unknown sample spaces, Schnell (2018) engaged nine pairs of 11–13 year old students in intensive exploration of a known sample space at different sample sizes. The game Betting King involved rolling a 20-sided die where 7 sides were red, 5 were green, 5 were yellow, and 3 were blue, and was accompanied by an Excel simulation. Students could choose what color to bet on and how many dice throws would be in the game (between 1 and 10,000). Normatively, the best strategy was to bet on red and roll 10,000 times to exploit the Empirical Law of Large Numbers to ensure red’s victory, similar to some of the problems in Wagner (2006) and to the polling problem in Bar-Hillel (1979).

Although Schnell (2018) was published after this dissertation’s data was collected, it is worth mentioning for several helpful features that could have supported students’ mechanistic reasoning. First, the Betting King context appeared to successfully motivate students to manipulate sample size as a causal factor to help them win the game. At the same time, the encouragement to explore the known sample space provides opportunities for students to see the relationship between sample space and what they are seeing for large samples. Although understanding among students varied, one student articulated the connection well: “[Red wins more] only for high numbers. Because it has more red sides” (Schnell, 2018, p. 5). And another student pair had the exchange: “A: [Red] is quite fast. B: Yes, maybe because there are many reds” (Schnell, 2018, p. 5).

However, most of these experiences fail to provide a reason why and how one would expect larger samples to be reliable. They essentially provide memorable contexts for observing that it does happen and provide situations for observing and exploiting the Empirical Law of Large Numbers. If these types of experiences are truly the foundation of the size-confidence intuition as Piaget and Inhelder (1951) suggest, the fragility of the size-confidence intuition demonstrated in many studies above may be because students lack a compelling account of the inter-level mechanistic relationships that explain why this law would hold. Schnell (2018) appears to succeed in promoting causal, if not mechanistic, reasoning through the use of a clever task that drives students to manipulating sample size and examining the sample space. The mechanism is still hidden, however, and the limited representations (bar graphs of how much each color wins) may not facilitate making the connections between frequency and sampling distributions or perhaps even applying these insights to sampling distributions at all. Students here see the operation of variability only on individual trials (Pratt et al., 2008), and can sometimes be coaxed into seeing the long-run stability at the emergent aggregate level (Ireland & Watson, 2009), but coordinating these two levels is challenging (cf. Stroup & Wilensky, 2014).

The theoretical analysis of Chi and colleagues (2012) suggests the location of the difficulty. The link between the local and aggregate level is dependent on what they term the *collective summing* mechanism, here the process by which each sample's values contribute to the shape of the sampling distribution. In their study of students' understanding of diffusion after receiving training in emergence and a computer simulation that showed both the macro and micro levels, the authors claim that their study "revealed that understanding this collective summing mechanism is crucial in generating correct causal explanations for more complicated science processes" (p. 51). Although Chi and colleagues do not present any empirical data to support this assertion, the fact that the collective summing mechanism is the point of linkage between the micro and macro levels certainly makes collective summing the most plausible crucial point for inter-level causal explanation. In the case of the sampling distribution, the collective summing mechanism is the mean.

2.3.2 Swamping and balancing: Student understanding of the mean

An empirical sampling distribution of the mean is itself composed of means. It is because of properties of the mean that these empirical sampling distributions of the mean have less variability with larger sample sizes. This is not true of all statistics, such as the sum. Below is reviewed the literature on students' understanding of the mean in hopes that the explorations of the conceptions of mean can illuminate the inter-level causal link between the single-sample properties of the mean and the relationship between mean, within-group variability, and sample size.

There is a long strain of literature documenting students' difficulty with the mean. Students tend to know the mean algorithm but are not able to understand what the mean represents. In Pollatsek, Lima, and Well's (1981) foundational study, they found that college students often blindly applied the algorithm to problems that involved weighted means or in figuring out the missing value for a dataset with a given mean. The authors surmised that few students had what they called *analog knowledge* of the mean, a visual or kinesthetic understanding of the mean in terms of the dataset. They point to the balance beam as one possible model that could support analog knowledge of the mean, since the mean can be viewed as the *center of balance* of a set of values. If values of a dataset were placed as weights on corresponding position on an ideal balance beam, with the fulcrum is placed at the mean, then the balance will be achieved. This is because the torque at any given point is the weight times the distance from the fulcrum, and the beam balances when all the torques are equal (e.g. the sum of the deviations from the fulcrum equals zero). The mean is the point at which the sum of the deviations from the mean is zero, so it must be the place where the fulcrum can be placed and the balance is kept in rotational equilibrium (see Marnich, 2008).

However, the pedagogical utility of the balance beam model is complicated by the frequent finding that students do not tend to have strong understandings of the balance beam (Hardiman, Well, & Pollatsek, 1984; Siegler, 1976). Fortunately, Hardiman and colleagues (1984) also provided evidence that understanding of the balance beam can be trained at least in college students. The authors first gave students an assessment that included several weighted mean problems and balancing problems modeled after Siegler (1976). Students with poor performance on the balancing problems were randomly assigned to a training group, which received an extensive training regime with a physical balance beam without mentioning the mean, or a control group, which received unrelated probability problems. On the posttest, a balance beam was given and participants were given a simple demonstration with two weights that the arithmetic mean could be thought of as the balancing point, and then were given several more weighted mean problems. The students who had undergone the training performed substantially better on the posttest than their own pretest and also better than the control students on weighted means problems, and referred to "balancing" strategies in their answers.

Center of balance is not the only pedagogical model for the mean, however. Another model is the *fair share* model—the mean is what would happen if the total number is shared equally among all cases. Strauss and Bichler (1988) identified seven properties of the mean and created a series of fair share problems—without explicitly identifying the arithmetic mean—to 8–14-year-old Israeli students. Of these properties, most older children understood that the mean-as-fair-share was located between the extremes of the data, was influenced by the addition of any number besides for the mean, did not necessarily appear in the data, and can be a non-integer even for discrete quantities (e.g. 1.6 children). However, the participants tended to perform poorly on the following problem:

> Children brought cookies to a party they were having. Some children brought many and some brought few. The children who brought many gave some to those who brought few until everyone had the same number of cookies. Was the number of cookies given by those who brought many the same as the number of cookies received by those who brought few? Was it more? Less? Why do you think so? (Strauss & Bichler, 1988)

This item was intended to measure whether students recognized that deviations from the mean necessarily sum to zero. The wording of the problem, however, permits reasoning correctly about this without necessarily understanding this property: Of those who correctly responded to the problem, many simply stated that since it was the same total number of cookies, the surplus of those who had more has to equal the deficit of those who had less. However, many of the correct answers at older ages involved stating the property of the means that the sum of the deviations has to equal zero. This study shows the possibility of representing this crucial feature of the mean within the fair share model.

Strauss and Bichler (1988) also attempted to analyze the representative nature of the mean, but their task simply asked, "For a class party, Ruth brought 5 pieces of candy, Yael brought 10 pieces, Nadav brought 20, and Ami brought 25. Can you tell me in one number how many pieces of candy each child brought?" There is no reason that the mean is the appropriate response here, and this is a rather impoverished notion of representativeness of the mean as a summary with no contextual motivation. Students mostly argued from contextual reasons and a concern that there be enough candy for everyone without really considering the arithmetic mean.

In their qualitative study of 4th, 6th, and 8th grade US students, Mokros and Russell (1995) also probed students' understandings of the arithmetic mean with an eye towards the concept of the mean's representativeness. The authors used a weighted means problem from Pollatsek and colleagues (1981), but introduced *construction tasks* where students created datasets to have a given mean, as well as an open-ended *interpretation task* where students interpreted a messy, bimodal dataset to determine the typical value and the highest amount (not necessarily the arithmetic mean) that could be considered typical. They found five levels of understanding of "average": mode, algorithm, reasonable, midpoint, and mathematical point of balance. Students who simply found the mode or mindlessly applied the algorithm did not display evidence of having any sense of an average as representative. Students who viewed the average as some kind of *reasonable* value, or as the *midpoint* (i.e. median) of the data had some idea of representativeness but had trouble coping with asymmetry. The two students with the highest conception—*mathematical point of balance*—had a more complex understanding of balance, although they had misconceptions about the ways in which balancing works. The authors argue that the conceptual understanding of the arithmetic mean should be taught before the formula, which seems to hijack students' understanding of the mean as potentially representative.

In their review of the literature on understandings of centers and averages, Konold and Pollatsek (2002) take this argument further by arguing that the representativeness of the mean relies to some extent on the "interpretation" of the mean. Konold and Pollatsek regard the Strauss and Bichler (1988) question to "tell me in one number" as *data reduction* and of little use in comparing groups or viewing the mean in a motivated statistical context. Although closer to some idea of central tendency, the authors contend that the *typical value* interpretation (Mokros & Russell, 1995) still does not suggest one of the most important features of measures of central tendency—the superiority of larger samples. Although the *fair share* interpretation (e.g., Strauss & Bichler, 1988) affords richer interpretations of the mean, Konold and Pollatsek (2002) argue that many situations make the reallocation principle of this interpretation rather abstract (since variables such as IQ cannot actually be shared, and yet we are interested in representing IQ). Moreover, they argue that the fair-share concept of average discourages examination of distribution and variability, since it is principled on the equal division of the total amount.

Rather, they argue that the most promising interpretation of the average is as a *signal in a noisy process*. The data is representing some sort of underlying process, and the average is estimating that signal. They view this as the most essentially statistical understanding that incorporates the true interpretations we want students to understand as well as the superiority of larger samples, since they will afford capturing more of the signal. This conception is most clear in three cases: 1) repeated measurements (signal = estimate of true mean, noise = error), 2) production (signal = desired target, noise = error) and 3) comparing groups (signal = group tendencies, noise = within-group variability). The promise of these contexts for developing understanding of the mean has been validated in later studies (e.g., Konold & Harradine, 2014; Lehrer & Schauble, 2007).

Konold and Pollatsek's (2002) schema of interpretation blurs two aspects of understanding of the mean: the *interpretation* within a context, and the *model* of the mean itself. Fair share, for example, can be either a model of the mechanism of taking the mean, or an interpretation of the resulting number. For instance, Groth (2005) provided a repeated measurements context—supposedly ideal for eliciting the signal-in-noise interpretation of the mean—but this was presented in a case-value plot that many participants solved by borrowing from the longer bars to “give” to the unknown bar using a fair-share model (equally distributing the total). Marnich (2008) misleadingly refers to the signal-in-noise interpretation as a *kind* of fair share model, arguing that the signal is “leveling out” the contributions of the noise.

The two primary models are the fair share analog and the center of balance. The fair share analog shows that one can physically find the mean of different values by redistributing the values until each one has the same height. In the center of balance model, the mean is the point at which the data set balances; the mean can be found from a stacked dotplot by making a series of “balancing moves” (always keeping the balance constant by moving the value of one case up one unit whenever moving the value of another case down one unit) until all the data values are piled up at the mean (e.g., O’Dell, 2012).

Marnich (2008) studied the relationship between the two models among college students and presented the mean to students randomly assigned to work through online trainings in fair share models, center of balance models, or a control training on problem-solving strategies. The students took an identical pretest and posttest that included problems intended to assess fair share understandings of the mean, center of balance understandings of the mean, and several mathematical principals of the mean. Like many other researchers, Marnich found many rote and inappropriate applications of the arithmetic mean formula. However, he found a promising ability for people to transfer the mathematical principal that the sum of the deviations from the mean is zero (one of the seven properties examined by Strauss & Bichler, 1988) across contexts intended to elicit the fair share model and those intended to elicit the center of balance model. He views this as a promising way of connecting the two models, and that people will have a better understanding of the mean if they are able to flexibly utilize both the fair share and the center of balance models. Moreover, he found that the strong understanding of fair share models may be an easier place for many students to start, especially because its link with the algorithm is straightforward, and because of the aforementioned difficulties with students' understanding of balance beams (e.g., Siegler, 1976). Marnich's proposed remedy is to show the sum of the deviations equaling zero in both models simultaneously to encourage students to develop stronger connections between the two models.

Overall, the literature on student understanding of the mean shows substantial challenges for building the desired inter-level causal understanding of the sampling distribution. As a collective summing mechanism, the mean operates relatively opaquely for students of all ages. Many conceive the mean as simply an algorithm and lack a flexible analogical understanding that sampling variability interventions would be able to exploit. Moreover, literature on student understanding of the mean does not directly address the issues of swamping and balancing at different sample sizes, two promising inter-level mechanistic explanations of the mean.

Perhaps one lesson that can be learned from the literature is that the manner of representing and modeling the data influences the way students are likely to conceptualize the mean, since fair-share models generally depend on a case-value plot and center of balance models depend on a stacked dotplot. Representations that support *swamping* could be created by directly representing sample size on the axis of a plot and showing how, as the sample size grows, the mean moves less and less as sample size increases. It is not clear what representations could more clearly support *balancing*, because the concept of balancing is itself rather challenging for students and students struggle with understanding how balance beams actually work.

2.3.3 Heaping

Although swamping does operate as a mechanism for the Empirical Law of Large Numbers, by itself swamping only explains the decreasing size of movements of the mean relative to a deviation and does not directly explain the fact that the sample mean converges to the population mean. For example, it is quite possible for a process to have decreasing possible movements without converging to any particular value: The harmonic series (1 + ½ + ⅓ + ⋯) has smaller and smaller steps, but never converges.

A more adequate explanation of the convergence of the sample mean to the population mean is here termed *heaping*: Viewed from the Laplacian perspective of a sample space with equally likely outcomes, there is an increasing proportion of possible outcomes near μ in the sampling distribution as sample size increases. This mechanism is inspired by research into the Blink Game (Konold & Kazak, 2008), where students judge the fairness of a game in which students draw either an open eye, "•" or a closed eye, "-" twice to make a face; they are then asked whether *stare* ("•,•"), *wink* ("•, -" or "-,•"), or *blink* ("-,-") are more likely, and students are guided to the understanding that wink is composed of two possible simple outcomes. In other words, there are more ways to get a *wink* than a *stare* and therefore the *wink* player will tend to win more frequently. Figure 2.2 extends this logic to the sampling distribution of a Bernoulli random variable with $p = 0.5$, which models a coin flip game where heads are scored as 1 and tails scored as 0, and the average score is recorded. When the sample size is 2, there are two ways to get an average score of 0.5. As the sample size increases, *heaping* can be observed. There are increasingly more ways of getting an average score close to population mean compared with farther away, and therefore sample means increasingly "heap" around the true mean.

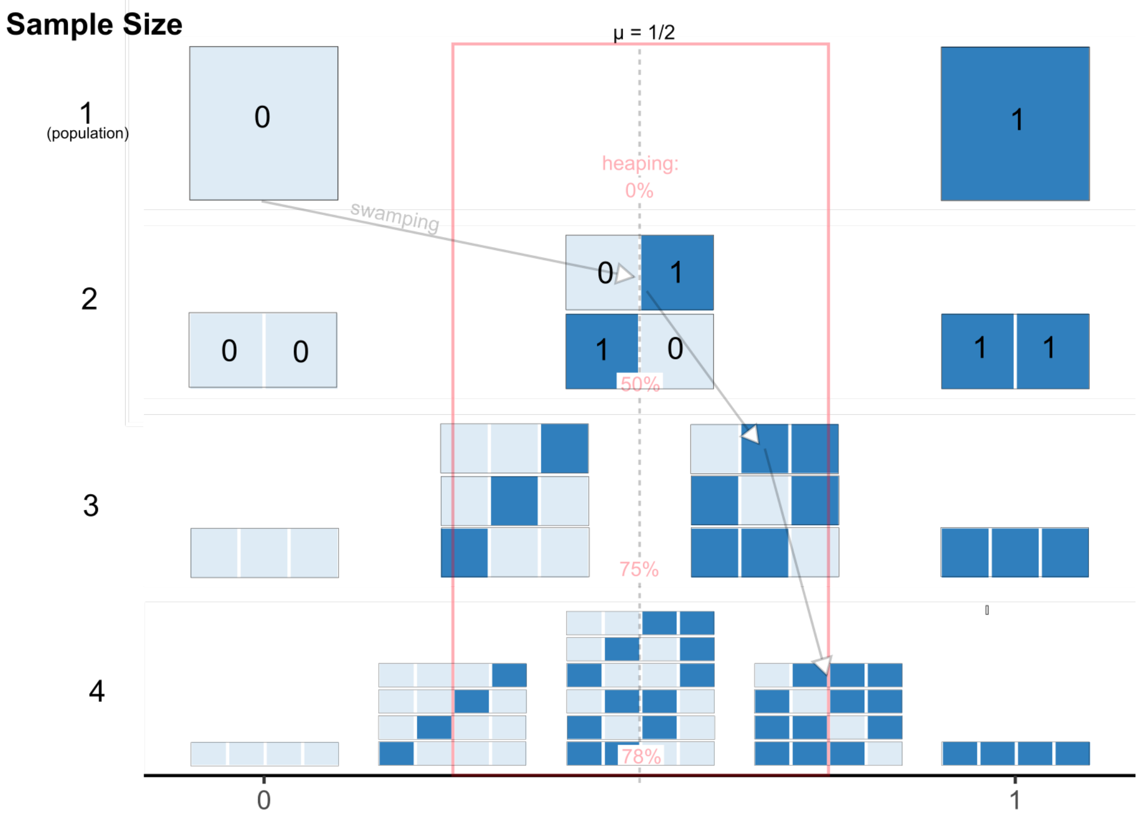


Figure 2.2. Illustrations of sampling distributions of the means of a Bernoulli random variable with $p = ½$ for sample sizes 1–4. Permutations of outcomes are shown that lead to a given result; e.g. for sample size 2, there are two ways to get a mean of 0.5: 0,1 or 1,0. Swamping is shown by the arrows: for adding successive 1s to an initial 0, the change in means becomes smaller and smaller, since the each additional 1 has decreasing effect as sample size grows. Heaping is shown by the increasing percentage of possible outcomes within 0.25 of the mean.

2.4 Conclusion and Research Questions

This literature review argued that existing research into students' understanding of sample size and sampling variability, while demonstrating persistent difficulties and enumerating some of the conditions for those difficulties, nevertheless did not clearly support any single existing theory for accounting for students' inconsistent explanations and decisions. Instead, a plurality of conceptions and decision-making processes were identified. Conceptual change is supported by developing a causal explanation for a phenomenon, perhaps particularly a mechanistic causal explanation that accounts for *how* the phenomenon emerges, yet prior interventions have had limited success and appear to provide quite limited support for mechanistic understanding. Three mechanisms that may be able to support mechanistic understanding are the size-confidence intuition, swamping, and heaping, but new research should ensure that representations are created that allow these mechanisms to become more clearly visible.

The next chapter of this dissertation describes a series of tasks, Growing Certain, that builds on the existing literature by providing representations intended to support students in reasoning mechanistically by supporting understanding of swamping and heaping through visualization. This study contributes to the literature by answering the following questions:

1. *How can we describe conceptual change when it happens in the context of Growing Certain? How do the changes observed relate to the instructional design, including the representations used, prompts, and social interactions?*
2. *How and when are students reasoning mechanistically before, during, and after Growing Certain? How does each element of Growing Certain (prompts, representations, etc.) interact with students' mechanistic reasoning about the Empirical Law of Large Numbers?*

The next chapter describes in more detail how Growing Certain was implemented and how the study was designed to answer the research questions.

Chapter 3

Methodology

This chapter presents the methodology for research investigating how students engage with causal reasoning about the Empirical Law of Large Numbers (Freudenthal, 1972; Sedlmeier & Gigerenzer, 1997), the principle that a larger sample size leads to lower sampling variability of the mean. The research is centered around a sequence of activities, called "Growing Certain", which attempts to provide opportunities for students to develop a *mechanistic causal explanation* for sampling variability. A mechanistic causal explanation explains the *process* that connects the input, sample size, with its caused output, sampling variability (Russ et al., 2008). Growing Certain targets two causal mechanisms for sampling variability as sample size increases: *swamping*, the decreasing influence of an extreme observations on the sample mean (Well et al., 1990); and what this dissertation has termed *heaping*, the increasing concentration of sample means around the true mean.

The Growing Certain activities are relatively novel and little is known about how students might reason in these circumstances. Therefore, this study had two exploratory research questions:

1. *How and when are students reasoning mechanistically before, during, and after Growing Certain? How does each element of Growing Certain (prompts, representations, etc.) interact with students' mechanistic reasoning about the Empirical Law of Large Numbers?*

2. *How can we describe conceptual change when it happens in the context of Growing Certain? How do the changes observed relate to the task design, including the representations used, prompts, and social interactions?*

Since the research questions were focused on students' individual sense-making processes, one-on-one clinical interviews were chosen as the data collection paradigm. Five students in a randomization-based introductory statistics course, CATALST (Garfield et al., 2012), attended five videorecorded semi-structured interviews with a clinical interviewer. In addition to prompting the students to explain their reasoning with both prewritten and improvised questions, the interviewer provided technical support with various software and logistics. However, the interviewer never explicitly taught about mechanistic understanding of sampling variability in order to avoid contaminating the students' sense-making process (diSessa, 2007). The research questions were originally intended to be answered answered by two complementary qualitative analyses of videorecordings and artifacts for one study participant: *mechanistic coding* (Russ et al., 2008) to capture student's mechanistic reasoning and *microgenetic coding* (Siegler & Crowley, 1991) to characterize the shifts and changes in students' thinking. The original analysis plan differed from the actual analysis, presented in the next chapter, but is presented here for transparency.

The next section summarizes the causal mechanisms of sampling variability and the accompanying representations that Growing Certain targeted. The remaining sections present the participants, materials, procedure, and analysis plan.

3.1 Targeted causal mechanisms and representations

Growing Certain primarily targets two causal mechanisms for the decrease in sampling variability as sample size increases: *swamping*, the decreasing influence of each value on the sample mean, and *heaping*, the increasing concentration of the sampling distribution around the true mean. Swamping was represented by a *sample size plot* of the mean against the sample size, while heaping was represented by visualizing the *permutations* that could produce each possible mean.

3.1.1 Swamping and the sample size plot

Growing Certain supports swamping through focusing student reasoning on what is here termed the *sample size plot.* This plot is of a single sample growing from sample size 1 to 500 with sample size represented by the *y* axis and sample mean value by the *x* axis (Figure 3.1). As more and more values are added to the sample, the mean goes from jumping around to gradually moving less and less while remaining close to the true mean. Although no research on these plots were found in the literature review, prior teaching documents have provided examples of this type of plot (e.g., DeCarli, 2012).

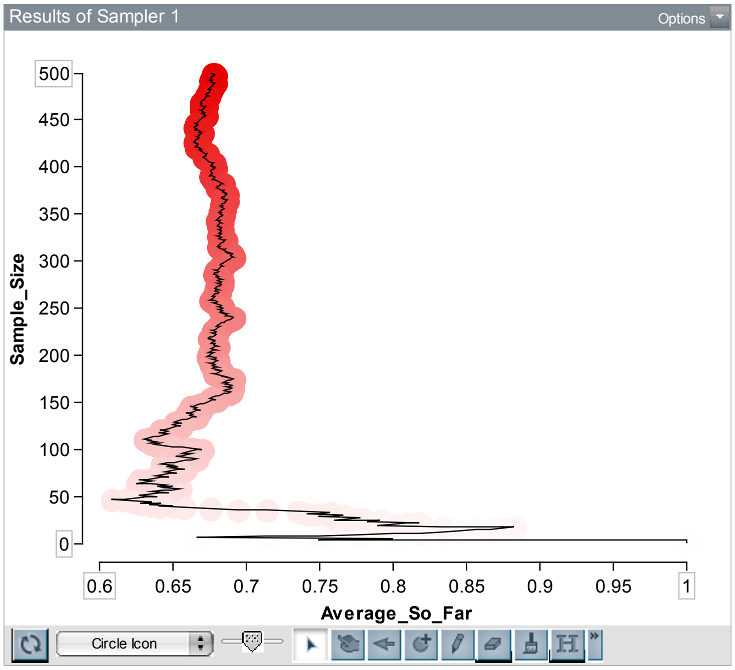


Figure 3.1. Sample size plot of a sample of 500 from a Bernoulli random variable with *p* = ⅔, created by a pilot participant with guidance from the researcher. The *x* axis shows the mean of the sample as it grows from 1 to 500.

This plot was hypothesized to support swamping because the contribution of a given value visibly decreases as the sample size increases. In Figure 3.1, above, each success of a Bernoulli random variable can be observed to change the mean less and less as the sample size grows. Connecting line segments between successive points are drawn to emphasize the value-to-value changes. In DeCarli (2012), the teacher drew a reference line on a similar plot at the true mean to focus students on the increasing closeness of the sample mean to the true mean. In contrast, Growing Certain excluded this reference line to focus students more on the changes in the mean as the sample size grows and thus to focus the representation more on the mechanism of swamping. Growing Certain also put mean on the *x* axis, as opposed to putting sample size on the *x* axis as in DeCarli (2012), for consistency with the empirical sampling distribution dotplots (which represent sample mean values on the *x* axis) that students use in the CATALST curriculum (Garfield et al., 2012; Zieffler & Catalysts for Change, 2017).

3.1.2 Heaping and permutations plots

Growing Certain supports reasoning about heaping by several plots that display the possible permutations of sample values that produce a given sample mean, inspired by the Blink Game (Konold & Kazak, 2008). After viewing empirical sampling distributions of a Bernoulli variable, students are shown how to use TinkerPlots™ to color the sample means by the different permutations that could produce that mean (Figure 3.2). This representation is hypothesized to help show the mechanism of heaping. The increasing number of permutations that produce sample means near the true mean value creates the increasing concentration of density near to the true mean value in the empirical sampling distributions. Growing Certain starts with empirical sampling distributions because these are familiar to students from the CATALST curriculum (Garfield et al., 2012; Zieffler & Catalysts for Change, 2017).

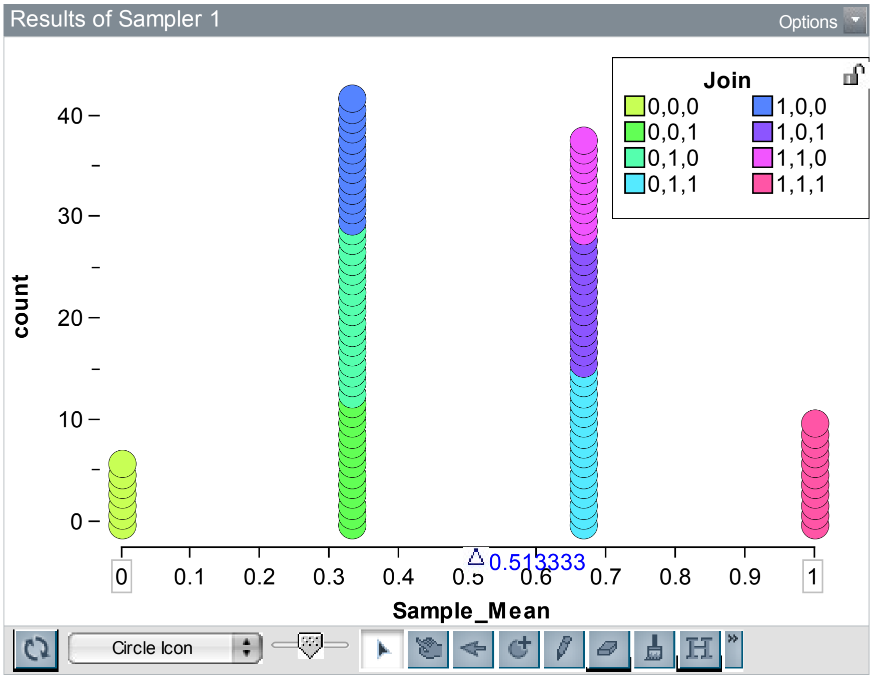


Figure 3.2. Empirical sampling distribution at *n* = 3 for the mean of a Bernoulli random variable with *p* = ½, colored by the permutation. Note that 0 and 1 only have one permutation, whereas ⅓ and ⅔ each have 3 permutations, leading to more mass of the empirical sampling distribution at those values.

Growing Certain then more deeply probes the sample space and the role of permutations by having students directly generate the different permutations that result in each mean using building blocks, inspired by the Family Game in Konold & Kazak (2008), which itself was inspired by the focus on permutations in Abrahamson (2006). Students construct sampling distributions for an equiprobable Bernoulli random variable by generating all possible permutations of white (0) and black (1) blocks at *n* = 2, *n = 3*, and *n* = 4 (Figure 3.3). Unlike Abrahamson (2006) and Konold & Kazak (2008), students explicitly compare samples of different sizes and place them on the same scale of the mean, whereas prior studies had students place samples by the number of a given color. The purpose of these prior explorations was primarily to support students in seeing the overall shape, whereas Growing Certain was particularly interested in supporting students' appreciation of the change in variability as sample size increased.

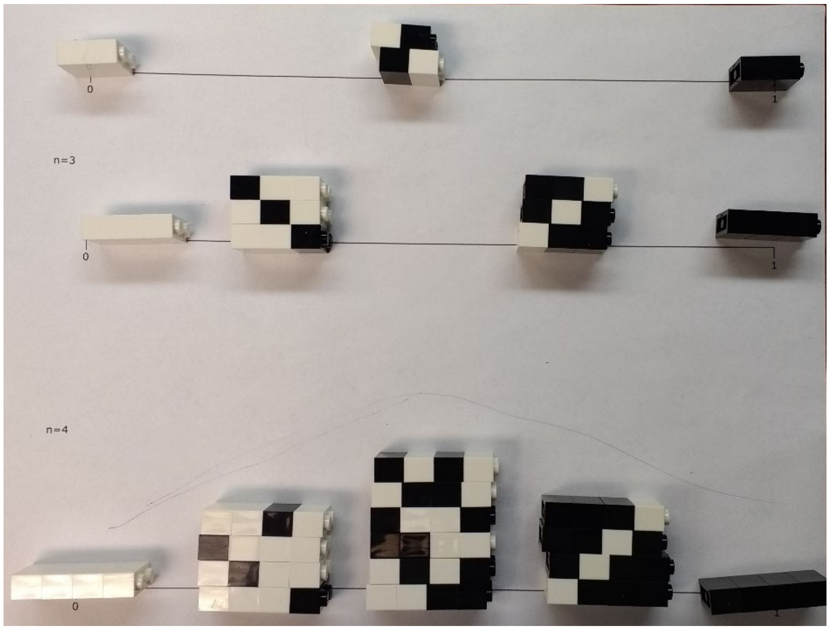


Figure 3.3. Theoretical sampling distributions of the mean of a Bernoulli random variable with p = ½ at n = 2 (top), n = *3* (middle), and n = 4 (bottom), created by a pilot participant using building blocks. Within a stack of blocks, each row of blocks represents a unique permutation of possible values—e.g., the black-black-white-black row represents the permutation 1-1-0-1.

3.1.3 Rationale for not targeting other candidate mechanisms

Growing Certain does not target three other student explanations that have appeared in the literature because they either did not fit well with the growing a sample paradigm, or they did not seem sufficiently strong as mechanistic causal explanations. Much prior attention in the Empirical Law of Large Numbers has centered around the *size-confidence intuition* (Sedlmeier, 1999) that larger samples are more like the population, and so are more likely to have a mean near the population mean (Well et al., 1990). Swamping and heaping can be demonstrated in the growing a sample context by the decreasing movement of sample means, or increasing proportion of outcomes near the population means, on an incremental level, even when comparing *n* with *n* + 1. However, it is not clear how adding a single extra value makes a sample "more similar" to the population, especially given that substantial variability in shape for small samples is already an instructional barrier (Konold & Kazak, 2008). Furthermore, the size-confidence intuition may be difficult for students to connect to empirical sampling distributions (Brown, 2015; Sedlmeier, 1999; Sedlmeier & Gigerenzer, 1997). A similar student explanation is *balancing*, the observation that larger samples provide more opportunity for large and small scores to balance out (Well et al., 1990). While balancing may be salient at two very different sample sizes, it is again not clear how one additional value leads to more opportunities for scores to balance out. If indeed the plausibility of size-confidence intuition and balancing depend on comparing two samples with very different sizes, then their applicability will be dependent on students' perception of the magnitude of the difference in sample sizes rather than simply an understanding that any difference in sample size is important as implied by swamping and heaping. Prior research has indicated sensitivity of sample size judgement to the difference in size of the samples (Lem et al., 2011; Murray et al., 1987; Obrecht et al., 2010).

Another possible target explanation is that larger samples capture a larger proportion of the population (Bar-Hillel, 1979). For finite populations, this explanation is true, provides a mechanistic causal explanation, and can be applied incrementally for comparing *n* and *n* + 1. However, sample-population ratio only practically applies when the sample is large relative to the population, and is irrelevant to infinite populations such as coin flips and dice rolls. Since the purpose of Growing Certain is to provide opportunities to develop a strong, mechanistically causal explanation that applies to a range of situations, bringing in the complications of finite populations may provide little leverage for building the strong understanding that students can transfer to a range of contexts. This speculation is not tested directly in the present study in order to avoid confusion around the role of the proportion of the population that prior research has uncovered (Bar-Hillel, 1979). However, both the pre- and post-interviews include items with finite populations and students' reasoning was examined to see how they appeared to use reasoning around the proportion of the population.

Growing Certain also did not target the property that the sample dotplot or histogram increasingly resembles the population distribution. Many prior interventions for the Empirical Law of Large Numbers target this property (e.g., Chance et al., 2004; Konold & Kazak, 2008). However, peoples' use of the similarity of sample to population can give rise to inaccurate heuristics and they may not always know when the similarity should normatively apply, which may lead to more confusion (Bar-Hillel, 1979; Kahneman & Tversky, 1972; Lem, 2015). In addition, causally explaining the resemblance of the entire distribution is inherently more complex than explaining only the properties of the mean. These speculations about the pedagogical limitations of sample representativeness were not tested directly in the present study. However, participants may have used this reasoning, and this study indirectly assessed how sample representativeness appears to support or hinder students' causal understanding of sampling variability by examining and probing participants. A more direct assessment, not attempted here, would be to compare Growing Certain with and without explicit scaffolding for sample representativeness.

3.2 Participants

The five study participants were students at the University of Minnesota enrolled in an introductory statistics course using the CATALST curriculum (Garfield et al., 2012; Zieffler & Catalysts for Change, 2017). Additionally, the five pilot participants were students who had completed the same CATALST course during the previous semester (see 3.3.1 Pilot testing, below). The CATALST curriculum emphasizes the logic of statistical inference through modeling and simulation. Throughout the course, students generate empirical sampling distributions using Monte Carlo simulation in the dynamically visual TinkerPlots™ software environment instead of using analytic normal-based formulas. Students first learn about modeling random processes, extend this to modeling the variability due to sampling, examine internal and external validity evidence using randomization and bootstrap tests, and conclude by estimating uncertainty using bootstrap confidence intervals.

CATALST students provided a useful testing ground for in-depth exploration of the Empirical Law of Large Numbers in an environment where students already had some of the basic conceptual tools to support their understanding. Many documented student difficulties with the Empirical Law of Large Numbers appear to stem from more fundamental misunderstandings about distributions, sampling, and sampling distributions (Chance et al., 2004). Since nearly every CATALST activity involves generating and interpreting empirical sampling distributions, studying CATALST students permits an exploration of reasoning about sampling variability when students already have extensive experience with sampling distributions. Furthermore, these students also have experience setting up and evaluating stochastic models in dynamic statistical software, TinkerPlots™. Finally, CATALST students are less likely to have encountered formula-based approaches to teaching statistics, since these are avoided (Garfield et al., 2012). Focusing students on algorithms may shortcut their development of conceptual mathematical understanding (Mokros & Russell, 1995). Students in this curriculum, however, have still struggled with understanding the relationship of sample size and sampling variability, so there are still opportunities for instruction (Brown, 2015).

For all these reasons, the design of Growing Certain was targeted towards CATALST students who had completed Unit 1 (Modeling and Simulation) and had progressed through at least half of Unit 2 (Modeling Sampling Variation), which includes an informal introduction to hypothesis testing. At this point in the course, students should have been able to simulate empirical sampling distributions in TinkerPlots™ , having done so in several activities and assignments in Unit 1. All five pilot participants, who were recent course completers, were able to successfully do so. Additionally, students were introduced to evaluating statistical hypotheses in the first several activities in Unit 2. Growing Certain makes no further assumptions on students' technical or statistical knowledge. However, it should be acknowledged that Growing Certain may still have been ambitious for these students, involving interacting with a number of novel representations and using TinkerPlots™ in new ways.

Students (for both the pilot test and the study) were recruited by emails sent to the classroom instructors who forwarded them on to their classes (see Appendix A: Correspondence with EPSY 3264 students, p. 314). Students were compensated with $15 cash per 1.25 hour session, and if they completed all 5 sessions they were given a $25 completion bonus for a total compensation of $100 for 6.25 hours. The cash incentive was meant to motivate not only the highest-performing students but also lower-performing students; in a prior study of this same population (while performing pilot testing for Brown, 2015), it was anecdotally observed that low monetary incentives tended to only motivate high-performing students. Students were accepted on a first-come, first-serve basis, since it was found in pilot testing that it was challenging to get students to commit to all five sessions. Consenting to participate in the study required consenting to be videorecorded, but students could choose whether or not their videos could be retained indefinitely at shown at professional meetings and conferences (Appendix A5: Consent form given to all participants, p. 316).

The design only required that one person participate in the study, consistent with prior detailed qualitative microgenetic analyses of learning (e.g., Ben-Zvi, 2006; Wagner, 2006). However, more students allowed a greater variety of experiences with Growing Certain to be examined to increase the dependability of the findings (Creswell, 2012). In the end, five students were able to participate in all five sessions; one participant only attended the first session. Only one student is analyzed in this dissertation (see Chapter 4), but the other four students with complete data could be analyzed in follow-up research. No demographic information was collected or analyzed, since the intent was to publish the full transcripts publicly as specified on the consent form. All participants are referred to in the results using the singular "they/their" since their preferred pronouns are not known.

3.3 Materials

The study consisted of a series of interviews, structured around the Growing Certain series of tasks. Table 3.1 shows the sequence of activities that provided opportunities for students to display and shift their understanding of sample size. The remainder of this section consists of detailed descriptons of these activities. More details about the activities are also available in Appendix B: Interview Protocols, p. 321; Appendix C: Pre- and post-interview questions, p. 337; and Appendix D: Complete Transcript for Participant S, p. 343. Activities with the same number (e.g. 3a and 3b) took place at the same interview time (e.g. Session 3). Students' reasoning was first explored in several sample size problems

(Activity 1a), and a replication of a classic demonstration of the Empirical Law of Large Numbers in the Post Office simulation (Well, Pollatsek, & Boyce, 1990; Activity 1b). Then, students participated in the six activities that comprised the Growing Certain sequence (Activities 2a–4b). Students explored growing a sample proportion from a dichotomous population (Activity 2a), then extending this to means and reasoning about the sample size plot (Activity 2b). These activities provided opportunities for students to envision the growing sample process and to describe the effect of sample size on sampling variability. Students then explored and simulated an open-ended problem with an unknown population model, in order to reason about the effects of sample size on inference (Activity 3a). These single-sample activities were designed to draw students' attention to features of the sample size plot intended to support reasoning about swamping. In Activity 3a, students also started viewing the empirical sampling distribution at different sample sizes, and for different unknown populations. Activity 3b built upon this shift as students were introduced to growing multiple samples simultaneously and viewing the permutations that comprised each possible sample mean. Students then constructed theoretical sampling distributions of discrete populations using concrete building blocks (Activity 4a) in a way intended to support reasoning about heaping and permutations. Afterwards, students explored the connections between the sample size plot and animations of simulated sampling distributions growing (Activity 4b), with the ability to view both plots of single samples and linked representations of multiple sample means simultaneously. In a post-interview (Activity 5), students responded to five additional sample size tasks, revisited their answers to the pre-interview questions, and provided concluding thoughts on their perceived learning. This final session provided an opportunity to see how students' responses to classic sample size tasks changed after participating in the Growing Certain sequence.

Table 3.1
Sequence of tasks

#	Title	Supported Goals	Representations
1a	Pre-Interview		Six textual problems
1b	Post Office Simulation		Dotplots of population, sample, and ESD
2a	Growing a Sample Proportion	Introduce “growing a sample” Recognize $\lvert p_n - p_{n-1}\rvert \to 0$ Recognize $p \to \pi$	Single sample with proportion annotation
2b	Growing a Sample Mean	Recognize $\lvert \mathrm{M}_\mathrm{n} - \mathrm{M}_\mathrm{n-1}\rvert \to 0$ Recognize $M \to \mu$ Swamping	Single sample with mean indicated Mean by sample size plot
3a	The Mystery Mean	Swamping Effect of *n* on ESD shape Inferential consequences	Hidden sampler contents Single sample with mean Mean by sample size plot, with divider
3b	Growing More Means	Introduce growing multiple samples Sample space Introduce heaping	ESDs, showing permutations
4a	Growing Possibilities	Heaping as combinatorics Process causes of $Var(M) \to 0$, $M \to \mu$ Connection between sample size plot and permutations plot	Permutations plots built of physical building blocks Sample size plots
4b	Growing Many Means	Explore heaping Revisit $Var(M) \to 0$ Revisit $M \to \mu$ Inferential consequences Regression to the mean Connections between sample size plot and empirical sampling distributions	Draw = 100 Sliders for sample number and sample size Plot of 200 sample means, explore with divider Single sample with mean Mean by sample size plot
5	Post-Interview		Five new textual problems and revisiting the five problems from the pre-interview

All representations are within TinkerPlots™ 2.3.1 (Konold & Miller, 2017) except for Growing Possibilities.
n = sample size; *p* = sample proportion; π = population proportion; →: approaches as *n* increases; *M* = sample mean; μ = population mean; ESD = Empirical Sampling Distribution

Most of the activities took place within TinkerPlots™ 2.3.1 (Konold & Miller, 2017). TinkerPlots™ allows visualizing how a random process is generated and interactively exploring simulation data. The ability to explicitly represent features of sampling distributions, building on the work of Chance et al. (2004), may have allowed students to see connections and develop a deeper mechanistic causal understanding of sampling variability.

3.3.1 Pilot testing

An initial pilot test was conducted of five students who had recently completed the CATALST course during the previous semester (Brown, 2018). The purpose of the pilot test was 1) to assess how richly the sequence of activities and prompts revealed students' thinking about sample size, 2) to assess whether the activities appeared to be successful in encouraging *causal* thinking about sampling variability, and 3) to improve the tasks, representations, and prompts to maximize 1) and 2). Various refinements were made to the tasks and representations throughout and after the pilot testing to focus the tasks and representations. These changes are discussed below in the context of each of the activities.

3.3.2 Activity 1a: Pre-Interview

Session 1 was primarily intended as a pre-interview to assess students' sample size reasoning on classic sample size problems before participating in the activities in Sessions 2, 3, and 4. After an introduction to the think-aloud process, students read each question, explained their reasoning and responses with prompting from the interviewer, wrote a few bullet points to summarize their reasoning, described how confident they were in the response, and then were asked to restate the problem that they had just solved once the interviewer removed the piece of paper.

Table 3.2
Pre-interview questions

Name	Variable type	Distribution type	Region of focus	Source
Hospital	Dichotomous	Sampling	Tail	Kahneman & Tversky (1972)
Referendum	Dichotomous	Frequency	Accuracy	Sabbag & Zieffler (2015)
Candy	Dichotomous	Sampling	Entire	Sabbag & Zieffler (2015)
Batting Average	Dichotomous	Sampling	Accuracy	Fong, Krantz, & Nisbett (1986)
Post Office	Continuous	Sampling	Tail	Well, Pollatsek, & Boyce (1990)
Casino	Discrete	--	--	--

Note. "Sampling Distribution" items focused on distributions of sample means, while "Frequency Distribution" items focused on individual sample values. "Tail" items focused on a tail of the distribution, while "accuracy" items focused on how close an observed mean/proportion was to the theoretical expectation and "entire" simply displayed the whole distribution.

The pre-interview questions are described in Table 3.2 and the full text of the questions is provided in Appendix C (p. 337). The problems were intended to provide some variety in exploring participants' understanding of sample size. The Hospital problem is a classic and frequently used sample size problem asking for comparison of which hospital, one with 15 or 45 babies born per day, is more likely to have more than 60% babies born per day (Kahneman & Tversky, 1972). The Referendum problem (Sabbag & Zieffler, 2015) asked whether a biased sample of 10,000 out of a population of 500,000 is sufficient for making generalizations, simultaneously assessing how students attend to bias in samples as well as their attention to the sample's proportion of the population (Bar-Hillel, 1979) as the primary factor in determining whether generalization is warranted. The Candy problem (Sabbag & Zieffler, 2015) was similar to the Hospital problem for comparing the graphs of distributions of proportions for a 50-50 variable at sample sizes of 10 and 100. The Batting Average problem (Fong, Krantz, & Nisbett,1986) assessed whether participants could recognize regression to the mean, while the Post Office problem was isomorphic to the Hospital problem with a continuous variable.

The final question, the Casino problem, was a new open-ended problem intended to get students thinking about the connection between sample size and inference:

> You work for the state casino regulation committee. Your job is to ensure that casinos are accurately reporting to customers the average winnings from slot machines. Suppose one slot machine pays out $0, $1, or $20 on each game, and the machine claims that the average payout is $0.90. You can play the slot machine as many times as you want, but it costs money each time. Construct a proposed strategy for determining whether the slot machine's claim is accurate.

This problem served several purposes. By emphasizing that the slot machine cost money each time to play, the problem provided a cue that unlimited sample size was not an option; the interviewer probed participants to pick a number of games to play to make an adequate determination. The problem also provided a context in which the superiority of larger samples begs explanation because of the cost of sample size, inviting students to express causal reasoning about sample size before being introduced to swamping and heaping in the explanation. Additionally, this question provided a general window into students' thinking about the process of statistical investigation.

Because most items were well-established, they did not undergo much revision during the pilot testing. The casino problem, however, was extensively revised, since initial versions of the item did not make it sufficiently clear that participants could play the slot machine, that there was a cost to playing the slot machine, or that the statistic of interest was the average payout. Additionally, the average payout in the problem was changed from $1.00 to $0.90, to help make clearer that this was an average payout that was not possible to gain from any single play.

3.3.3 Activity 1b: Post Office Simulation

After working through the text problems in the pre-interview, students explored a simulation of the Post Office problem in TinkerPlots™ to mirror the simulations originally explored by Well et al. (1990). Participants were given a dataset of heights and weights sampled from a dataset simulated based on historical Hong Kong data (Dinov, 2017). Students plotted the population, then created samples of size 10, then created the sampling distribution of means of the size 10 samples. This task was similar to prior tasks they had done in class (Zieffler & Catalysts for Change, 2017). Before actually displaying the sampling distributions of size 10, the interviewer ensured that students understood the different pieces of the simulation and were asked what they predicted the sampling distribution would look like; this was repeated for sampling distributions of size 100. As in the original Well et al. (1990) experiment, this assessed how students understood sampling distributions when they had substantial visual support for their thinking and were less likely to confuse the sampling distribution with the frequency distribution. Additionally, the approach of demonstrating the effects of sample size without forming a causal explanation mirrored intervention approaches found in prior literature (Chance et al., 2004; Sedlmeier, 1999; Well et al., 1990), and allowed exploration of how these participants functioned based on this existing approach.

During pilot testing, one participant did not notice that the TinkerPlots™ simulation results diverged from their expectation that the empirical sampling distributions of 10 and 100 would be similar because of TinkerPlots™ automatically rescaling them. Therefore, when participants were prompted for their predicted graphs, they drew them all on the same scale of 62 to 74, and the researcher intervened by adjusting the TinkerPlots™ scale to match the scale on the predicted graphs (Figure 3.4).

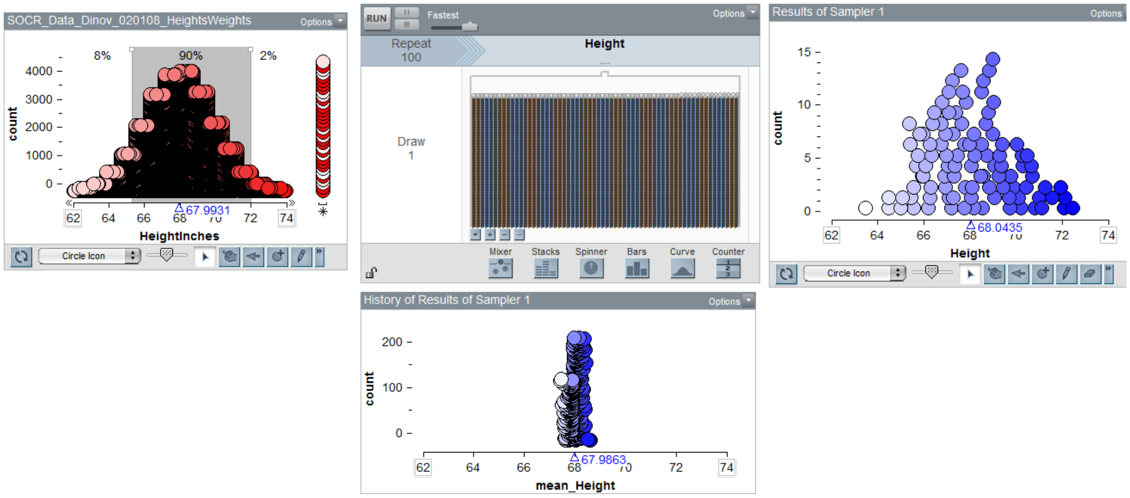
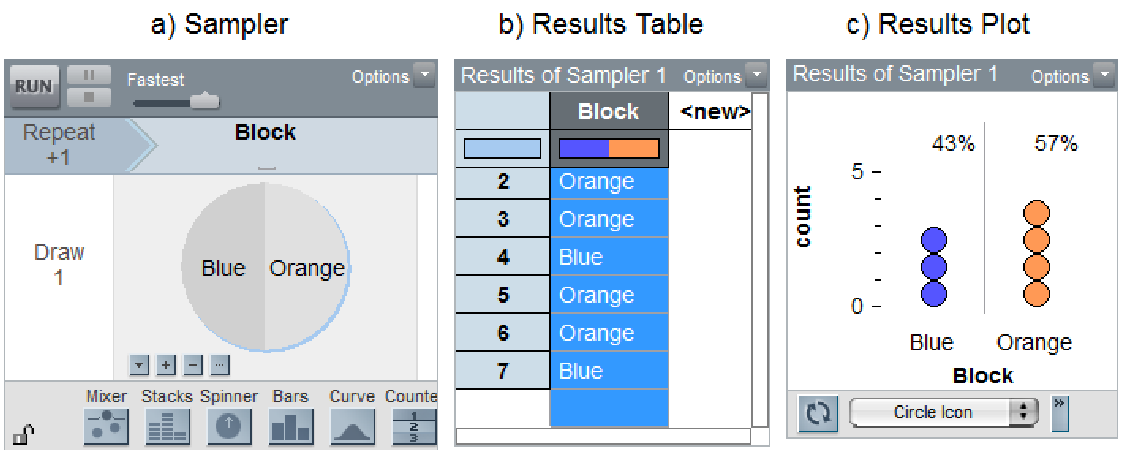


Figure 3.4. TinkerPlots™ setup for Post Office Simulation. Upper left, dotplot of the height of the entire hypothetical town; upper middle, mixer containing one ball for each of the town's heights; upper right, a single sample of 100 heights; bottom, empirical sampling distribution of means of 100 heights.

3.3.4 Activity 2a: Growing Sample Proportions

This activity familiarized students with the process of "growing a sample" (Bakker, 2004) for a dichotomous variable. Students first tracked the proportion of blue blocks while the researcher physically drew ten blocks with replacement from a box containing one blue and one orange block. Students made predictions about how they expected the percentage of blue blocks to change as the sample size grows, and manually kept track of the proportion they saw at each sample size. Then, students were introduced to the sample size plot (DeCarli, 2012) in TinkerPlots™ using dummy data, which had the sample proportion (or later, mean) on the *x*-axis and the sample size on the *y*-axis. Students were asked to predict what they expected to happen as the sample size grew for the blue and orange blocks by drawing the sample size plot that they would expect.

Then, students were asked to set up a model of the blue and orange blocks in TinkerPlots™ based on what they learned in class in activities such as the Introduction to Monte Carlo Simulation activity (Zieffler & Catalysts for Change, 2017, pp. 23–36). Students were then given guidance to on how to set up TinkerPlots™ so that they could repeatedly click the `RUN` button to add onto an existing sample (Figure 3.5). Questions prompted students to notice how much the proportion changes for each individual value added. Students grew a sample to size 50 and informally watched the proportion change, graphing the pattern that that they observed in a sample size plot.


a) Sampler b) Results Table c) Results Plot

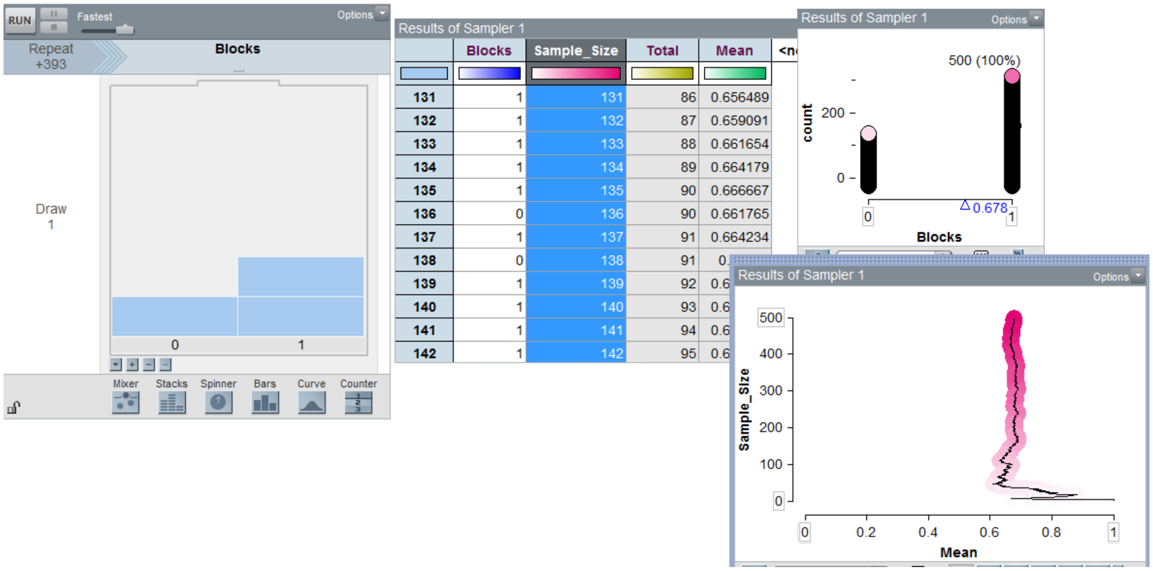
	Block	<new>
2	Orange	
3	Orange	
4	Blue	
5	Orange	
6	Orange	
7	Blue	

Figure 3.5. TinkerPlots™ setup for Growing Sample Proportions activity. Students set up model in Sampler: here "Blue" and "Orange" can be drawn with equal probability. When students click Run in the Sampler (a), a value is added to the Results Table (b). Students could also create a Results Plot (c) that shows a stacked dotplot of the number of heads and tails with the proportions for each.

This activity was deliberately kept simple to give students an opportunity to track the different pieces of growing a sample, using a simple context and a similar TinkerPlots™ setup to what they have already done in the course. This setup was also quite similar to several prior growing a sample activities that have been done with middle school students (Bakker, 2004; Lee et al., 2010; Pratt et al., 2008).

3.3.5 Activity 2b: Growing a Sample Mean

This activity extended Growing a Sample Proportion to the mean of a discrete distribution as well as showing linked representations in TinkerPlots™ of the sample size plot. Still using the context of the blue and orange blocks, students were told that blue blocks would be scored as 1 and orange blocks would be scored as 0. Students updated their simulation and grew a sample, now calculating the mean instead of the proportion. With the interviewer's help, the students' file was set up so that the mean at each sample size was calculated, and students were prompted to create the sample size plot in TinkerPlots™ so that they could view this plot growing as sample size increases. This procedure was repeated with a situation where there were *two* blue blocks and one orange block in the box, in order to extend beyond equiprobable situations where students' reasoning may operate differently (Figure 3.6).



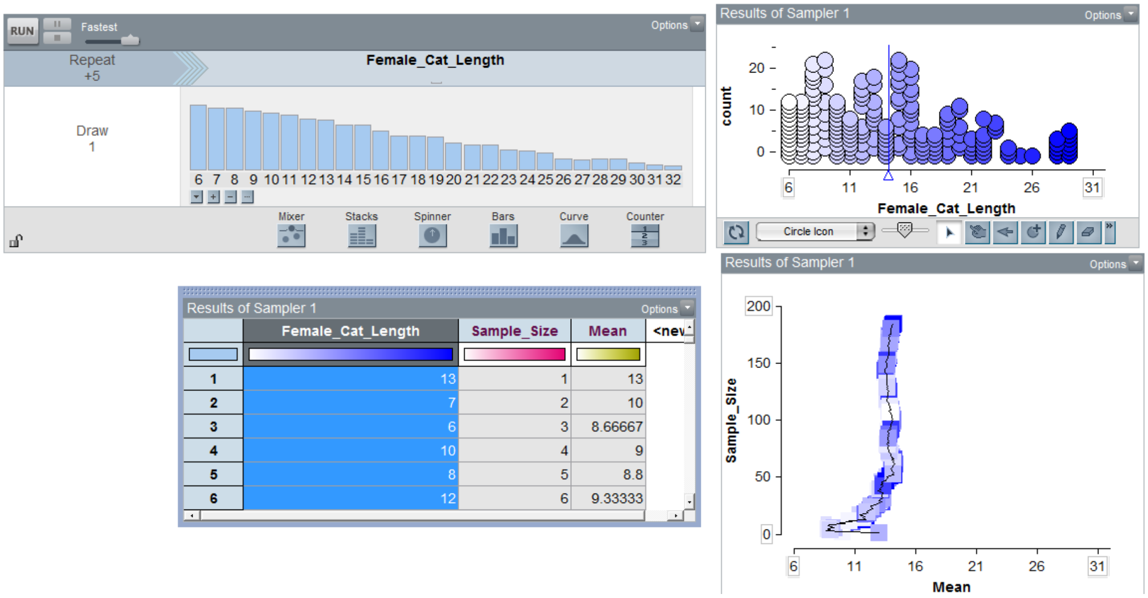
Results of Sampler 1

	Blocks	Sample_Size	Total	Mean	<n
131	1	131	86	0.656489	
132	1	132	87	0.659091	
133	1	133	88	0.661654	
134	1	134	89	0.664179	
135	1	135	90	0.666667	
136	0	136	90	0.661765	
137	1	137	91	0.664234	
138	0	138	91	0.	
139	1	139	92	0.6	
140	1	140	93	0.6	
141	1	141	94	0.6	
142	1	142	95	0.6	

Figure 3.6. Growing a Sample Mean TinkerPlots™ setup, created by a pilot participant and reformatted for clarity. Population model samples from a 0 and two 1s, left; table contains calculated fields to compute the total so far and mean so far; upper right shows dotplot of sample; lower right shows sample size plot (see text for details).

The availability of the sample size plot was a key feature of this activity, because it simultaneously provided a representation of sample size increasing and an assessment tool for eliciting students’ predictions about what they expected to happen. Students clicked a button to draw one more sample value and could view how this changed the mean both by looking at a plot of the sample and by looking at the sample size plot. Throughout the explorations, students were prompted to explain why they saw the patterns they did at different sample sizes, and attention was drawn to the spikes in the graph.

These representations were extended to more general distributions in the next task, which asked students to create a “cat factory”—a TinkerPlots™ sampler that generated cats of different lengths. The context was again familiar to students from the CATALST activity Generating Random Data—Cat Factory (Konold et al., 2007; Zieffler & Catalysts for Change, 2017, pp. 12–19). The Cat Factory context provided an opportunity to explore the relationship between distribution shape and the sample size plot. Students first drew a population shape that they felt was reasonable for a distribution of female cat lengths and observed the sample size plot for this distribution. It was expected that they would create a unimodal bell-shaped distribution, as all five pilot participants did. However, participants were then asked to draw a distribution that they felt would produce a different sample size plot and why, which provided a richer view of their understanding of the relationship between sample size, population shape, and sampling variability (Figure 3.7).



Results of Sampler 1

	Female_Cat_Length	Sample_Size	Mean	<new
1	13	1	13	
2	7	2	10	
3	6	3	8.66667	
4	10	4	9	
5	8	5	8.8	
6	12	6	9.33333	

Figure 3.7. Cat Factory TinkerPlots™ Setup. *Sampler*, upper left, shows a continuous distribution, drawn by a pilot participant to produce a different result than the unimodal distribution they drew initially; *Results Table*, lower left, calculates sample size and the cumulative mean of the current value and all previously drawn values; *Results Plot*, upper right, shows a dotplot of the sample and indicates the mean; *Sample Size Plot*, lower right, plots the means as sample size grows.

The main refinement that was made to the session during pilot testing was spending less time on the transition from the proportions to the means of a 0-1 variable. Initially in pilot testing, the interviewer emphasized that the mean of a 0-1 variable is the same as the proportion of 1s, but this assumption of a teacher role on the part of the interviewer seemed to disrupt the sense of the interview session as a way of exploring the sense-making of the participant (diSessa, 2007). Since the connection between proportions and means was not a major focus, this connection was deemphasized.

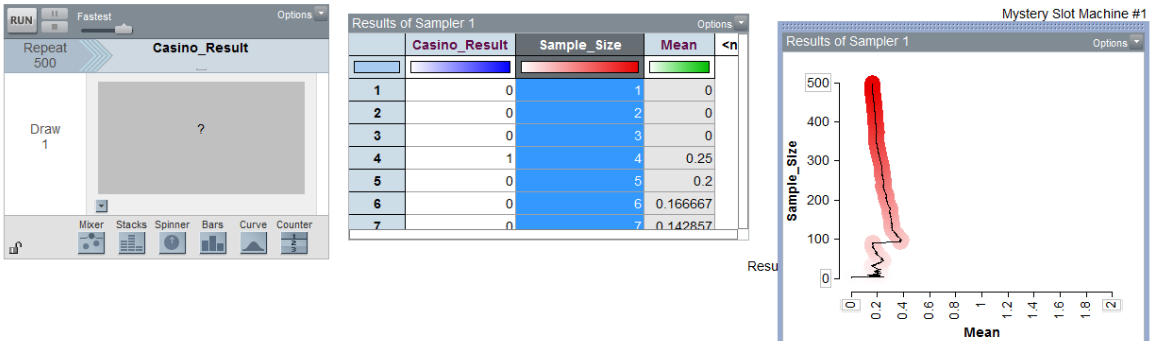
Additionally, one pilot participant did not display any swamping-like reasoning even after a fair amount of probing. Focused questions about the horizontal spikes in the sample size plot were included in order to give participants more opportunities to express this type of reasoning, since Well et al. (1990) found that swamping was more prevalent when participants were focused on the extremes of the distribution.

3.3.6 Activity 3a: The Mystery Mean

This activity was intended to assess students' understanding of the place of sample size within a statistical investigation. To more closely accommodate students' thinking, this activity was closer to a semi-structured interview than the relatively scripted Activities 1a–2b. The activity provided opportunities for students to apply and explore the Empirical Law of Large Numbers to the casino problem, introduced in the pre-interview (Activity 1a).

First, students were given the casino problem, and again decided on a strategy for how they would test whether the casino had accurately reported the average payout of the slot machine. The interviewer focused the conversation on sample size by probing students about how many games they would want to play on the machine.

Normatively, the sample size required to test the casino's claimed average payout depends both on the expected value and variability of the actual distribution. After students gave their response to the casino problem, they were given a sequence of two "mystery machines"—example simulations of how the machine *could* work. Each mystery machine consisted of a TinkerPlots™ file where the sampler (the population model) had hidden contents, each with possible values of 0, 1, or 20, consistent with the casino problem. This population was meant to make swamping salient, since the 20 will cause noticeable jumps in the mean plot but decreasingly so as the sample size grows. Mystery Machine #1 (Figure 3.8) had low dispersion (84.4% 0s, 15.2% 1s, .3% 20s) and an average (22¢) far from the claimed average of 90¢, while Mystery Machine #2 had greater dispersion (70.9% 0s, 27.6% 1s, 1.5% 20s) and a closer average (57¢). The second situation was intended to challenge students' assumption that small sample sizes such as 50 or 100 would be adequate for assessing the casino's claim.



Results of Sampler 1 Options

	Casino_Result	Sample_Size	Mean	<n
1	0	1	0	
2	0	2	0	
3	0	3	0	
4	1	4	0.25	
5	0	5	0.2	
6	0	6	0.166667	
7	0	7	0.142857	

Figure 3.8. The Mystery Mean TinkerPlots™ setup, with sampler setup hidden from students (indicated by "?"). Jumps in the sample size plot (right) are visible when a high value of 20 is drawn, but these jumps decrease in magnitude as the sample size grows.

For each Mystery Machine, students first sketched what they expected to see for the sample size plot assuming that the casino's claim was accurate, up until their chosen sample size. They then plotted the sample size plot from the TinkerPlots™ simulation and were asked whether they thought they had enough information to assess whether the machine's claim was accurate. Looking at the sample size plots was intended to reinforce attention toward the jumps in the graph, and then towards swamping.

Next, participants were asked to predict the empirical sampling distribution for their chosen sample size for the Mystery Machine, and compared this with the TinkerPlots™ simulation results. Students were expected to be familiar with generating the empirical sampling distribution in TinkerPlots™ since this was required in a typical CATALST course class activity (Zieffler & Catalysts for Change, 2017). The interviewer probed students about whether they thought the smaller sample size was adequate for assessing the casino's claim, and whether a different sample size would allow them to be more confident. Students then predicted the empirical sampling distribution and generated TinkerPlots™ simulation results for comparison. The comparison of different sampling distributions was intended to introduce students to viewing the sampling distribution as a way of comparing sampling variability at different sample sizes. Additionally, this activity served as a bridge between the single-sample focus of Activities 2a and 2b and the multiple-sample focus of Activities 3b–4b, and thus between the sample size plot and the permutations plot.

This activity developed into its final form during pilot testing. Initially, the session did not include the empirical sampling distributions, and only one mystery machine was presented. Two issues were noted for the first two pilot participants for this session: 1) these students appeared to struggle with the transition between growing a single sample and growing multiple samples, and 2) viewing only the sample size plot for one mystery machine did not seem to cause them to display any new reasoning about the casino problem. To address these issues, the interview began to probe more explicitly ahead of time about what sample size would be sufficient to evaluate the casino's claim of the $0.90 average payout and introduced the simulation of empirical sampling distributions. Additionally, introducing two mystery machines as hypothetical scenarios to help them think through the problem appeared to support richer reasoning about the context for later pilot participants.

3.3.7 Activity 3b: Growing More Means

Building on the re-introduction of empirical sampling distributions in Activity 3a, this activity had students grow multiple means simultaneously to draw attention to the *heaping* of outcomes around the mean. This was implemented in TinkerPlots™ by using the number of simultaneous draws in the *Sampler* as the sample size increases, and calculating the means of these samples. After being introduced to the basic setup for $n = 2$ of an equiprobable 0-1 dichotomous variable, students were asked to predict what they expected the empirical sampling distribution to look like, and then sketched what they actually observed in TinkerPlots. It was expected that many students would regard 0, 0.5, and 1 as equally likely outcomes based on the Blink Game, a similar activity for middle schoolers (Konold & Kazak, 2008), and several pilot participants made this error.

Students proceeded to go through the same process of predicting and modelling for $n = 3$ and 4. At this point, participants were introduced to viewing the possible permutations that made up each of the outcomes, by color-coding each of the means by the permutations of 0s and 1s that comprised it (see Figure 3.2, above). This was introduced alongside the empirical sampling distribution in order to give students an opportunity to connect the simulation results with the possible outcomes, key to the idea of heaping and explored more fully in Activity 4a, Growing Possibilities. The interviewer probed students about what they expected to happen as the sample size continued increasing.

This process was then repeated for a sampler which contained a 0, a 1, and an additional element, "1*". The element "1*" represented a second "1", but was labeled "1*" so that the student could track which "1" was included in each sample. The reason for including a separate 1 and 1* was to support students' attention to the different possible permutations in preparation for Activity 4a, since that activity emphasized that for $n = 2$, the outcome of 1-1* is different from 1-1 and 1*-1. Including a non-50-50 variable was, again, intended to help students generalize the principle of heaping beyond the context of a 50-50 variable.

In the original version of this activity shown to early pilot participants, students also saw an outcomes plot with only the *unique* permutations, and thus representing the full permutations plot (similar to Abrahamson, 2006). However, this essentially gave away the permutations that students would discover in the building blocks component of Activity 4a, and seemed to overwhelm the pilot participants with too many new representations. Since it was of interest how students reasoned about the permutations without being given them ahead of time, this view of the unique permutations was removed from this activity to allow a more gradual buildup to Activity 4a.

3.3.8 Activity 4a: Growing Possibilities

This activity allowed students to explore the phenomenon of heaping in more depth. Students created representations of sampling distributions with white building blocks representing a draw of 0 and black building blocks representing a draw of 1. First, students created the theoretical sampling distributions of all possible permutations for sample sizes of 1, 2, 3, and 4 (see Figure 3.3, above). This was then extended to a situation where there was one 0 and two 1s, with red representing one of the possible 1s, for sample sizes of 1, 2, and 3, to demonstrate the heaping effect for a non-symmetric variable (Figure 3.9).

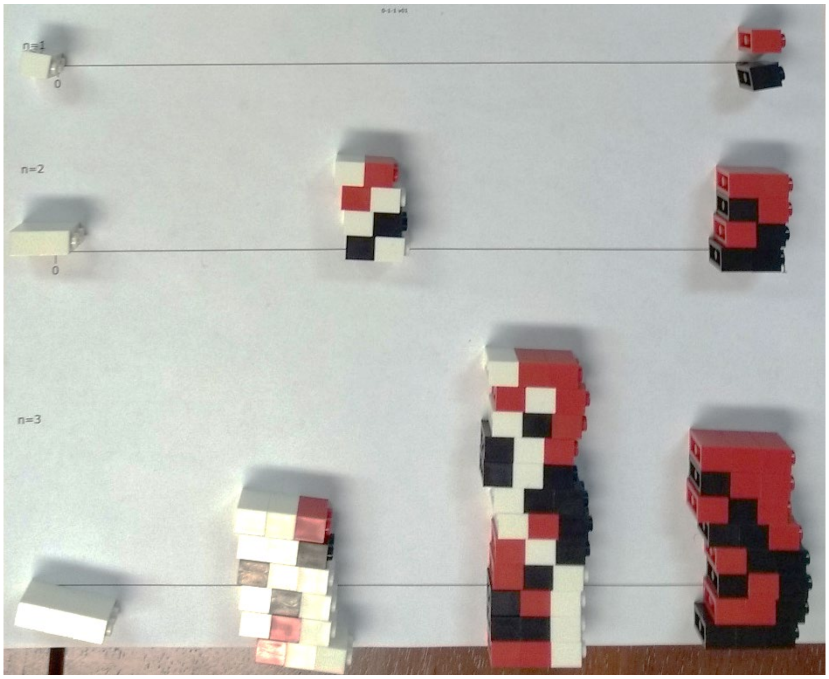


Figure 3.9. Sampling distributions of the mean of a Bernoulli random variable with *p* = ⅔ at *n* = 1 (top), *n* = 2 (middle), and *n* = 3 (bottom), created by a pilot participant using building blocks. Within a stack of blocks, each row of blocks represents a unique permutation of possible values—e.g., the white-red-black row represents the permutation 0-1-1.

In early pilot testing, this activity was loosely structured, with participants directed to construct the theoretical sampling distribution for each level and then questioned about what they noticed. However, participants seemed to be directing more of their attention to generating the permutations at each level than to the connections between levels. In later pilot testing and in the final study, the interviewer asked a structured series of questions that provided more opportunities for participants to attend to both heaping and swamping. In particular, participants were asked to first generate all the outcomes that ended with a white block (0) and to describe how these related to the outcomes found at the previous level (so all previous outcomes were represented, shifted over to the left), and similarly for black blocks (1s), along with red blocks (1s) for the white-black-red task. Participants were asked at each level what mean values they thought were most and least likely.

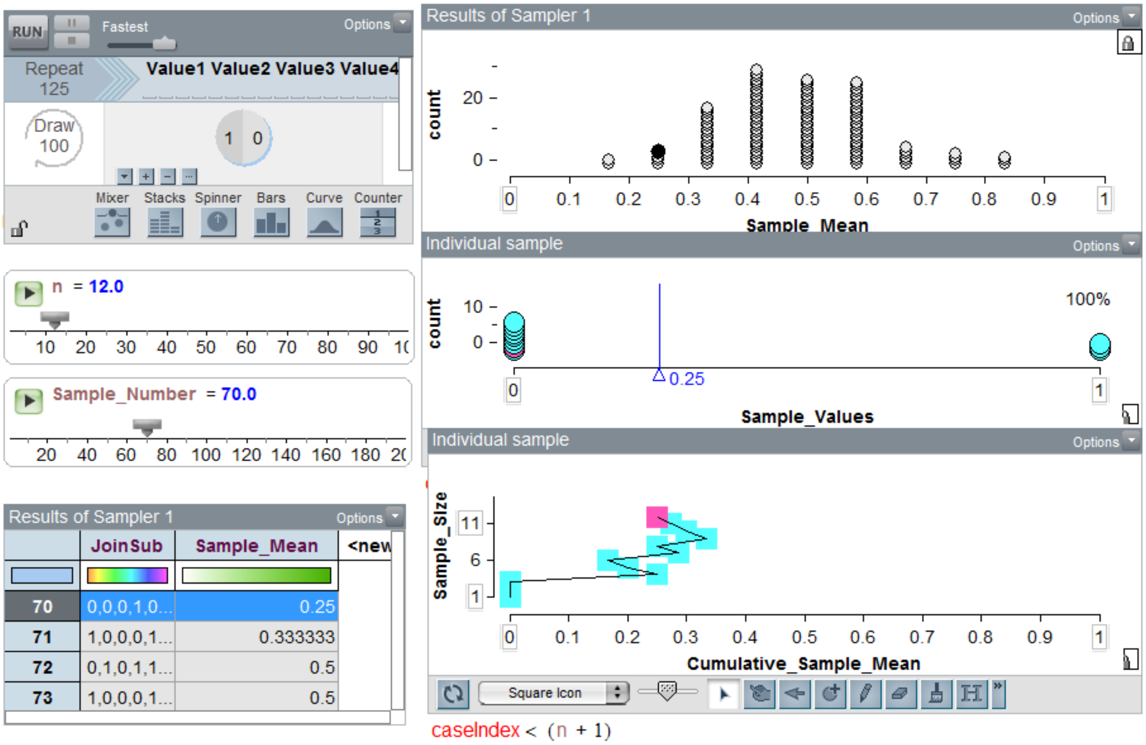
After generating all the theoretical sampling distributions for one of the populations, participants were asked what would happen if they kept going, and then were asked to plot the sample size plots for several different samples as represented by the blocks in the sampling distribution. For instance, if the interviewer pointed to a sample that consisted of the blocks black, white, black, black, their sample size plot would start at 1, go to 0.5, then 0.66, then 0.75. This segment was created to underscore the connection between the sampling distribution and the sample size plot so that participants could potentially see both heaping and swamping working simultaneously.

After participants worked through the white-black block task, the interviewer explicitly introduced an iterative strategy for generating and checking the theoretical sampling distribution from the theoretical sampling distribution of the previous sample size. This strategy involved recreating all the samples from the previous sample size, adding a white block, and shifting these to the left, and then doing the same for samples ending in a black block. This ensured that participants did not miss any combination, and moreover emphasized the connections between the outcomes at different sample sizes rather than simply generating the permutations for each sample size in isolation. The engine of heaping lies in how more outcomes end up piling up near the true mean, and so drawing students' attention to the connections between levels was conjectured to support heaping. The interview also emphasized these correspondences between levels in probes to provide more opportunities for students to recognize them and use them.

The introduction of this explicit strategy was after the students had already worked on the first equiprobable population to allow students to find their own strategies if they wished. Students were not required to use the interviewer's strategy after it was introduced to avoid students engaging with the task in an overly rote/mechanical way. No pilot participants appeared to use this strategy spontaneously, which may have led to their working memory mostly being taxed by trying to find permutations rather than seeing the emergent pattern of heaping.

3.3.9 Activity 4b: Growing Many Means

Students then interacted with a TinkerPlots™ simulation (Figure 3.10) that allowed them to manipulate sample size and view simulations of swamping (via the sample size plot) and heaping (via converging sampling distributions) simultaneously.



Repeat 125

Draw 100

n = 12.0

Sample_Number = 70.0

Results of Sampler 1

	JoinSub	Sample_Mean	<new
70	0,0,0,1,0...	0.25	
71	1,0,0,0,1...	0.333333	
72	0,1,0,1,1...	0.5	
73	1,0,0,0,1...	0.5	

caseIndex < (n + 1)

Figure 3.10. Growing Many Means TinkerPlots™ setup, with simultaneous representations of samples and sampling distributions. The *Sampler* is set up to simulate 125 samples of 100 coin flips. The ***n*** *Slider* determines what sample size is shown in the three plots; currently, samples of size 12 are shown. Each sample in the sampling distribution can be individually viewed: the ***Sample_Number*** *Slider* determines which sample is depicted in the plot of sample values (middle right) and the mean by sample size plot (lower right), with that same sample also highlighted in the empirical sampling distribution (upper right).

Students first viewed a TinkerPlots™ file only showing the sampling distribution just to introduce the new interface of controlling sample size with a TinkerPlots™ slider. After students briefly explored this, the interviewer revealed the plot of the individual sample and the sample size plot. Students could select any sample via the slider, which highlighted the sample in the empirical sampling distribution, with the individual sample composition showed in a window below, and then the sample size plot shown in a window at the bottom. All windows were aligned so that the mean values lined up across all three plots.

The interviewer then directed students back to the plots of samples that they had identified in Activity 4a. Students then found a sample in their current TinkerPlots™ simulation that matched the previous sample, and watched an animation of all samples growing from n = 4 to n = 25, attending both to the sample size plot and to the change in the sampling distribution. Students then watched as the sample grew to 100. The reason for drawing students' attention to samples from Activity 4a was to help make explicit the connections between the permutations and the growth of simulated samples.

Well, Pollatsek and Boyce (1990) reported that participants attended to swamping more when their attention was on extreme values, and initial pilot results suggested that this may also be true for heaping. Therefore, the interviewer then displayed only the most extreme sample means at n = 15 and asked the student to predict what they expected to happen to these extreme samples as more values were added to those samples. In the ensuing animation, the means of these extreme samples tended to migrate toward the overall population value. Again, this process was repeated for a variable where 1 was twice as likely as 0.

3.3.10 Activity 5: Post-Interview

The post-interview mirrored the pre-interview, with five different classic sample size problems for students to solve (Table 3.3) in a similar fashion to the pre-interview.

Students then were presented the pre-interview problems again and solved them. The interviewer then showed the student their original responses to the pre-interview problems, and asked how their understanding of the problem had changed since when they originally saw it. Students were then asked if they saw connections between the problems and the sample size plot or the permutations plot, what they thought the purpose of all the study tasks were, what they noticed about their thinking changing, and general comments about what they did or did not like about the activities.

Table 3.3
Post-interview questions

Name	Variable type	Distribution type	Region of focus	Source
Geology	Continuous	Sampling	Tail	Well et al. (1990)
Factory	Dichotomous	Sampling	Accuracy	Fong, Krantz, & Nisbett (1986)
Exam Preparation	Continuous	Sampling	Tail	Garfield, delMas, & Zieffler (2012); Brown (2015)
Coin Flips	Dichotomous	Sampling	Tail	Ziegler (2014)
Working Choices	Dichotomous	Frequency	Accuracy	Fong et al. (1986)

Note. "Sampling Distribution" items focused on distributions of sample means, while "Frequency Distribution" items focused on individual sample values. "Tail" items focused on a tail of the distribution, while "Accuracy" items focused on how close an observed mean/proportion was to the theoretical expectation and "Center" focused on the center of the distribution.

The Factory, Working Choices, and Exam Preparation questions all went through substantial revision during the pilot tests. Factory and Working Choices, both from Fong et al. (1986), were verbose problems with a lot of contextual information that threw pilot participants off, and were focused and clarified substantially as pilot testing progressed.

The Exam Preparation problem, which asked how the randomization *p*-value would change of a group comparison when the sample size increased, was further clarified after some pilot participants did not remember how randomization tests worked.

The post-interview was intended to provide some account of how participant's thinking about classic sample size problems changed, both in terms of reasoning and correctness. New problems were presented first to prevent contamination by the old responses, but the interviewer also presented the old problems to see how students' reasoning changed on those problems. Since the intervention was centered around two representations and causal explanations—the sample size plot for swamping and the permutations plot for heaping—the interviewer probed participants to see if they were consciously making these intended connections. Finally, more open questions explored participants' perceptions of the study and the statistical situations contained therein.

3.4 Procedure

Participants participated in the clinical interviews in a quiet lab setting. The lab room had one laptop with screen capture software along with a video camera to provide the level of detailed data on student behavior often considered necessary for microgenetic analyses of learning (see Analysis, below; Parnafes & diSessa, 2013). This researcher administered all interviews. As noted above, a teaching role was avoided in relationship to the mechanistic reasoning under study in order to create an environment in which students' authentic sense-making can be explored (diSessa, 2007). The researcher did step in to clarify activities, software issues, representations, and very occasionally declarative knowledge about statistics (e.g., one student asked about how to understand which direction a distribution was skewed), but never any principles or mechanisms relating to sample size. Instead, the researcher gave affirmative and positive responses to encourage the participant to feel comfortable sharing their thoughts in a natural conversational way, without regard to normative correctness (e.g., “Mm-hmm”, “OK, great”, “Sounds good”). Given the focus on mechanistic causal reasoning, the researcher frequently asked about why the student thought a particular response was true, and improvised many other probes to understand participants’ reasoning.

To prepare participants for the modality of the interview, participants were read a script encouraging them to share all their opinions and reactions, and emphasizing that the researcher was more interested in the participants’ reasoning than whether or not they gave a “correct” answer. Participants practiced this principle on two common think-aloud questions, asking how many windows there were in the place where they lived, and how difficult it was to get to the interview. The researcher modeled the same attitude of affirmative responses and probing reasoning on these practice problems. After participants answered these questions, the researcher again reminded them that even though they were going to answer some statistical questions, that he was not their statistics teacher who wanted a correct answer, but just a researcher interested in how they think about the problems. Participants then participated in the first interview (Activities 1a and 1b, above). After each interview, the participants were paid cash as noted above (see section 3.2, Participants). In subsequent interviews, the researcher began the interview immediately without any practice questions.

The researcher performed all transcription services. Screen capture videos were uploaded privately to YouTube, which added subtitles using automatic speech recognition. These automatic transcripts were corrected, reformatted to indicate speakers, interruptions, etc., and with added commentary to note on-screen activities. Screenshots from the screen capture video were interlaced through the transcript to clarify what participants were viewing at different points, which could differ based on the participant. Analysis ended up focusing on one focal participant whose transcript is provided in Appendix D (p. 343).

3.5 Analysis Plan

This section presents the original analysis plan before the data was coded. As described in the next chapter, the coding was tweaked in order to answer the research questions by developing a mechanistic grammar for the Empirical Law of Large Numbers (see section 4.1). However, the original plan is here presented for transparency into the research process, and because the eventual coding system addressed the research questions in a fundamentally similar way to the original plan and was fully consistent with the approach specified in the prospectus paper for this dissertation. The original plan was to answer the research questions using two complementary qualitative coding techniques. *Mechanistic coding* would have deductively coded student's mechanistic reasoning using a short *a priori* list of general codes (Russ et al., 2008), whereas the inductive *microgenetic coding* would stay close to the particular data to flag the shifts and changes in students' thinking. These coding techniques would be supplemented with examination of the raw transcripts, artifacts, and video in order to answer the research questions. The final coding system, specified in the next chapter, combined these two coding systems into a mechanistic "grammar" to closely track S's mechanistic reasoning at a fine-grained level.

3.5.1 Mechanistic Coding

The originally planned mechanistic coding was based on Russ et al. (2008), who developed a scheme for coding students' use of mechanistic reasoning while problem-solving (Table 3.4). The Russ et al. (2008) codes were developed based on theories from the philosophy of science about how professional scientists reason about mechanism (Machamer et al., 2000).

Table 3.4
Originally planned mechanistic discourse analysis codes, adapted from Russ et al. (2008)

Code	Description	Example
Describing the Target Phenomenon	Students clearly state or demonstrate the particular phenomenon or result they are trying to explain	"Larger and smaller samples will have the same variability."
Identifying the Setup Conditions	Students identify particular enabling conditions of the environment that allow the mechanism to run	"'One' is twice as likely as 'Zero' in the spinner."
Identifying Entities	Students recognize objects that affect the outcome of the phenomenon.	"I added a new value onto the sample."
Identifying Activities	Students articulate the actions and interactions that occur among entities and their properties.	"The new value in the sample changed the mean."
Identifying Properties of Entities	Students articulate general properties that are necessary for this particular mechanism to run.	"This one has sample size 500."
Identifying Organization of Entities	Students attend to how the entities are structured in relationship to other entities.	"The sample size is small relative to the population size"
Chaining: Backward and Forward	Students make claims about what must have happened previously to bring about the current state of things (backward) or what will happen next given that certain entities or activities are present now (forward).	"If you keep taking averages, the outliers are going to have less effect on your data, so it will get more and more narrower."
Analogies	Students compare the target phenomenon to another mechanism	"It's like flipping a coin – no matter how many times you flip it, it's always 50%."

Since the Russ et al. (2008) mechanistic coding was developed and refined on classroom interactions of 1st graders discussing physics, several adaptations were made to the differing context of the present study. First, the Russ et al. coding system included a category for "Animated Models", which included gestures indicating how different parts of the mechanism interact. These were not considered applicable to the abstract and emergent mechanism of sampling variability. Second, Russ et al. (2008) divided up their transcripts into chunks which they called “conversational turns”, perhaps to facilitate calculations of interrater reliability and coding in Microsoft Excel. However, data reduction of this sort is discouraged in microgenetic learning analysis due to the focus on a *small grain size* of analysis (Parnafes & diSessa, 2013). Since the present study had only one coder, coding did not involve explicitly breaking down transcripts into chunks in order to keep the analyses consistent.

Consistent with Russ et al. (2008), the same words of a transcript could have multiple simultaneous mechanistic codes. For instance, participants might note that an extreme value causes a large jump in the sample size plot. Under the original analysis plan, this would have been coded as Identifying Entities (an extreme value, a large jump in the sample size plot) and Identifying Activities (the fact that the extreme value caused the jump). Additionally, none of these codes have a normative implication: A section of the transcript where the participant says that large samples and small samples have the same variability would have been coded as Describing the Target Phenomenon even though it is an incorrect description.

3.5.2 Microgenetic Coding

Microgenetic analysis emphasizes observations of individuals throughout a period of change, with a high density of observations relative to the rate of change, and intensive trial-by-trial analyses to infer the processes giving rise to the change (Siegler & Crowley, 1991). This approach has been adapted to study conceptual change in a learning environment as well, with a focus on developing and improving theories of change and termed *microgenetic learning analysis* (Parnafes & diSessa, 2013). Parnafes and diSessa (2013) list several key features of data interpretation for microgenetic learning analysis: 1) keeping close to the data, 2) two levels of interpretations (constructing a literal interpretation before constructing a theoretical one), and 3) alternative interpretation, which involves constructing alternative accounts of students' thinking and pitting them against each other. Wagner (2006) successfully applied this approach in his detailed study of one students' learning the Empirical Law of Large Numbers, a study which provides some precedent for the current one. Microgenetic learning analysis is quite similar to interpretive microanalysis, which has been successfully applied in several student inquiry studies in statistics education (e.g., Ben-Zvi, 2006).

Because microgenetic learning analysis emphasizes the importance of staying close to the raw data without reduction (Parnafes & diSessa, 2013), the planned microgenetic coding employed as light a touch as possible. Within a framework of inductive open coding, several types of codes in the original analysis plan were constructed to support the microgenetic analysis: *explicit change, in vivo, causation. Explicit change* would be when a student indicated awareness that they were changing their mind, or that their prediction did not in fact come true. Such codes could be used examining the role of explicit cognitive conflict, a hallmark of classical conceptual change (Posner et al., 1982) and the basis of many previous interventions for teaching the Empirical Law of Large Numbers (e.g., Chance et al., 2004). *In Vivo* coding (e.g., Miles, Huberman, & Saldaña, 2014, p. 74) would have used words and phrases drawn directly from the student's transcript to flag an idea.

Although multiple snippets could be given the same In Vivo code, keeping the student's phrasing could be used to keep the focus on the student's particular way of expressing an idea and to draw attention to repeated phrases. *Causation* coding (e.g., Miles et al., 2014, p. 79) was intended to flag attributions about how and why particular outcomes occur,. Causation coding would involve mapping the actual connections between variables—so a student saying "small samples are more likely to vary" would be coded "Small N → Variability". Causation coding directly inspired the mechanistic grammar developed in the next chapter.

3.5.3 Answering Research Question 1

1. *How and when are students reasoning mechanistically before, during, and after Growing Certain? How does each element of Growing Certain (prompts, representations, etc.) interact with students' mechanistic reasoning about the Empirical Law of Large Numbers?*

Each students' mechanistic codes were examined to illuminate how students were reasoning mechanistically throughout the interviews. One finding of Russ et al. (2008) was that differing levels of complexity of codes tended to cluster throughout the analysis. For instance, the Russ et al. (2008) students tended to spend a lot of time early on Defining the Phenomenon, which was considered the least sophisticated engagement with mechanisms. Eventually they shifted to more depth by examining the Entities and Activities more closely, and finally transitioned to more sophisticated mechanistic understanding as displayed by Chaining when students were able to "play" the mechanism forward and backwards to reason about what had to be true in the past and make strong predictions about the future.

In the present study, the focal student's data were examined to determine how and when clustering may occur throughout the experiment. One potential progression is that students only displayed lower-level codes during the pre-interview, gained increasingly rich codes during the activities, and were able to sustain those high-level codes during the post-interview. Another possibility is that students gained increasing sophistication with swamping in the first half of the interviews, but then when confronted with sampling distributions reverted back to low-level codes until they gained more experience with permutations plots and heaping in Activities 3b–4b, at which point they eventually were able to use Chaining again in that context. Many other progressions were conceivable, since students did not come in as blank slates and already had intuitions and senses of mechanism around the Empirical Law of Large Numbers. Although the mechanistic grammar involved a different coding scheme, Chapter 5 presents the focal students' progression ithe eight high-level mechanistic categories proposed by Russ et al. (2008).

Answering this question also involved examining the relationship between the representations, prompts, and students' mechanistic reasoning. The representations viewed were the same across participants, but the researcher's prompts and probes sometimes differed due to the semi-structured nature of some of the interviews. Therefore, analysis also examined whether the researcher was prompting students about entities, activities, etc. For instance, there were only two problems throughout Growing Certain that mentioned population size, and these were the only problems where pilot participants attended to the relationship between sample size and population size. The prompts used may provide valuable information on what levels of the mechanism students were attending to.

3.5.4 Answering Research Question 2

2. *How can we describe conceptual change when it happens in the context of Growing Certain? How do the changes observed relate to the task design, including the representations used, prompts, and social interactions?*

For research question 2, the focal student's shifts and changes—both explicit and implicit—were holistically examined throughout the interviews. The microgenetic coding was intended to be the primary guide for constructing narratives for students' shifts, and the final analysis used a combination of the mechanistic grammar and direct examination of the transcripts and artifacts. Explicit shifts such as a recognition that a prediction did not come true were examined, but also implicit shifts such as subtly changing explanations in response to different contexts.

Additionally, this research question overlapped with the first research question, since it was expected that some of the shifts and changes would specifically be in regards to students' mechanistic reasoning. The interplay between the coding and evidence from the raw data was examined to see how these changes may relate to the task design and representations used.

Chapter 4

Results

This chapter presents the results of the Growing Certain problem-solving interviews described in the previous chapter. The analysis was designed to answer the following exploratory research questions:

1. *How and when are students reasoning mechanistically before, during, and after Growing Certain? How does each element of Growing Certain (prompts, representations, etc.) interact with students' mechanistic reasoning about the Empirical Law of Large Numbers?*
2. *How can we describe conceptual change when it happens in the context of Growing Certain? How do the changes observed relate to the task design, including the representations used, prompts, and social interactions?*

Because of the volume of data generated by the five study participants, this dissertation focuses on just one participant in order to describe and explore their mechanistic reasoning on a fine-grained level. "S" was chosen as the focal participant because "they" were genuinely engaged in sense-making throughout the activities, a key criteria for the ecological validity of clinical interviews (diSessa, 2007). S seemed very comfortable thinking aloud and expressing their opinions throughout the interviews, and did not seem to be trying to give an answer that the interviewer liked or understand; for instance, repeated probes about an unusual response did not seem to deter S from that opinion. S was also a useful participant because they did not seem especially reflective or sensitive to contradictions in their reasoning, which meant that they were comfortable stating these contradictions aloud without self-editing. Finally, S was also the last participant, which meant that fewer interviewer errors occurred during S's sessions than with other study participants.

Several typographic conventions are used below in the excerpts from the transcript. The interview and line numbers for each excerpt are given in parenthesis in the form (1.236–238), where "1" indicates that the quote is from the first interview and "236–238" indicates the range of line numbers. Short interjections by the other person are given in brackets (e.g., in the quote "…just like when you flip a coin. [OK] Um …" the interviewer said "OK" in the middle of S's speech). The interjection [mh] indicates any affirmative non-word sound (e.g. "Mm-hmm"), [Hm] any neutral non-word sound (e.g. "Huh" or "Hmm"), and [Uh-uh] any negative non-word sound. Other non-verbal notes are given in between forward slashes, e.g. "/laughs/". Very quiet speech, such as when S spoke aloud to themselves while figuring out what to do in TinkerPlots™, was surrounded by curly braces (e.g., "{Let's see here…}").

The coding procedure departed from the original design in order to be able to answer the research questions, and a "mechanistic grammar" was developed for describing reasoning about the Empirical Law of Large Numbers, described in the next section. The subsequent sections provide summaries of the coding and S's mechanistic reasoning during each of the five interviews: Pre-Interview, Growing Sample Proportions & Means, The

Mystery Mean & Growing More Means, Growing Possibilities & Many Means, and the Post-Interview

4.1 A Mechanistic Grammar for the Empirical Law of Large Numbers

Early on in the analysis, it became clear that just coding the eight high-level mechanistic categories proposed by Russ et al. (2008)—Describing the Target Phenomenon, Identifying the Setup Conditions, Identifying Entities, Identifying Activities, Identifying Properties of Entities, Identifying Organization of Entities, Chaining: Backward and Forward, and Analogies—was inadequate to richly capture S's mechanistic reasoning. For example, in S's responses to the Hospital problem, they used "chaining"—sophisticated reasoning about what will happen now given the state of the mechanism—but they conflated the sample size and the number of samples in the empirical sampling distribution (ESD). A lot would be missed by simply coding that segment "chaining", since students' reasoning can fluidly shift, incorporate new thoughts, and misclassify objects and their properties when examining an emergent phenomenon such as the Empirical Law of Large Numbers (Chi, 2013). Since the activities were all in the domain of sampling variability, it was usually possible to identify the components of the mechanism of sampling variability that S was referring to. As coding progressed, a lot of similar structural relationships were emerging, and freehand coding became unwieldy. Therefore, a symbolic "grammar" was developed to allow systematic description of S's attention to the mechanistic components described by Russ et al. (2008) and Machamer et al. (2000): *entities*, *properties*, *actions*, and *relationships*. *Entities* are the "things" or agents in the mechanism, such as when S attended to a *Sample* or a *Population*. Entities have characteristics or *properties* that influence the running of the mechanism, such as S's attention to *Sample size*. Entities and their properties may have *actions* or activities, such as how S noted that the *Sample mean varies* as the *Sample size increases*. *Relationships* are any interaction between entities, properties or activities that S verbalized. For instance, S noted that if a *Set case* [entity] *is subtracted* [activity], this **causes** [relationship] the *Set* [entity] *proportion* [property] *to decrease* [activity].

Table 4.1

Prevalence of Entity, Property, and Action Codes that Occurred in More than One Interview for Participant S

Entity	Property	Action	Prevalence by Interview 1 2 3 4 5
ESD			▄▁▅▃▄
ESD		centers	▂▁▂▁▁
ESD	center		▁▁▂▂▁
ESD	number of slots		▁▁▂▃▁
ESD	number of slots	increases	▁▁▂▃▁
ESD	proportion of upper extreme		▄▁▁▁▄
ESD	range		▃▁▂▁▂
ESD	range	decreases	▂▁▂▁▁
ESD	shape		▂▁▃▂▁
ESD	size		▄▁▂▁▂
ESD sample			▄▁▅▃▂
ESD sample	mean		▁▁▃▂▁
ESD sample	size		▄▁▅▃▂
ESD sample	size	increases	▂▁▄▃▁
Population			▅▆▆▅▄
Population	mean		▃▃▅▁▂
Population	proportion		▃▂▂▂▂
Population	values		▁▁▂▃▁
Population case			▂▂▄▂▁
Population case	probability		▁▂▂▁▁
Population case	value		▂▁▄▂▁
Set			▃▂▁▁▁
Set	proportion		▃▂▁▁▁
Set	proportion	changes	▃▂▁▁▁
Set	size		▃▂▁▁▁
Set case			▃▂▁▁▁
Set case		is subtracted	▂▂▁▁▁

Entity	Property	Action	Prevalence by Interview 1 2 3 4 5
Sample			█████
Sample	mean		▃▅██▄
Sample	mean	centers	▁▄▂▄▁
Sample	mean	changes	▂▃▄▄▂
Sample	mean	curves	▁▂▂▁▁
Sample	mean	decreases	▁▂▁▄▁
Sample	mean	increases	▁▁▂▃▁
Sample	mean	stays	▁▃▃▄▂
Sample	mean	varies	▂▄▃▁▂
Sample	mean	zigzags	▁▅▂▅▂
Sample	mean range		▁▂▂▁▁
Sample	number of slots		▁▂▁▃▁
Sample	population ratio		▂▁▁▁▂
Sample	probability		▁▁▃▄▂
Sample	proportion		▄▃▂▁▄
Sample	proportion	changes	▃▃▁▁▄
Sample	proportion	decreases	▂▃▁▁▁
Sample	proportion	varies	▂▂▁▁▃
Sample	sampling		▄▁▁▁▂
Sample	size		▆█▆▆▅
Sample	size	increases	▁█▅▅▁
Sample	streak		▁▃▁▂▂
Sample	total		▁▂▂▁▁
Sample	values		▁▁▃▂▂
Sample case			▄█▅▃▄
Sample case		is added	▃▄▂▃▁
Sample case		varies	▂▁▂▁▁
Sample case	probability		▁▁▂▂▁
Sample case	value		▂▃▄▂▂
Slot Machine			▂▁▃▁▂
Slot Machine	number of people		▂▁▂▁▂

Note. The bar height for Prevalence by Interview indicates the proportion of sections within an interview that contained S mentioning the relevant Entity, Property, or Action. ESD = Empirical Sampling Distribution.

Table 4.1 shows the full set of entities, properties, and actions which occurred in more than one interview. Several top-level entities related to sampling variability mechanisms that occurred in these data. The most abstract was *Set*, any collection of measures, which could be categorical or quantitative. Set properties that occurred in these data included *Set total*, the sum of the measures; *Set mean*, the arithmetic mean of the set; *Set proportion*, the proportion of successes (for dichotomous outcomes); and *Set size*, the number of elements in the set. Sets could also have *subentities*—elements that are somehow part of the set that are also entities themselves. An individual element of the set, for example, would be denoted as *Set case*. Most of the time, S discussed specific sets: *Sample*, a sample of values from a population or process; Empirical Sampling Distribution or *ESD*, the means of many samples. *Population* was often set-like, representing the values that were being sampled from, but since there was not any clear distinction in S's utterance between a population and process, it also acted in a process-like way sometimes, including properties such as *Population sampling is random*. The distinction between properties and entities was not a strict one. In very similar contexts S might be discussing the behavior of the mean of a sample as sample size grows, and depending on the emphasis and surrounding discussion this mean might be coded as *ESD sample mean* or simply *Sample mean*. With the activities' focus on the "growing a sample" paradigm, a common code here was *Sample size increases*, indicating a situation where S noted that the sample size was increasing.

In assigning these codes to S's utterances, the coding was intended to not run ahead of the understanding that the student actually showed. For example, in equiprobable situations, S repeatedly emphasized that the overall population chances were 50-50, and so it made sense to code *Population proportion* when S was talking about how the proportion would be close to 50% as the sample size increased. However, the student did not clearly connect the long-run average of a 0-1-1 variable with the *Population mean*, or to have a very clear sense about what this long-run average would be. Therefore, in this case there would not be any occurrence of the *Population mean* code, and instead there may be codes for *Population mode,* or if they simply thought it would be centered "higher" than a 0-1 variable, that would be coded *as Sample size increases, the Sample mean will probably be centered higher.*

This detailed mechanistic coding strategy replaced both the microgenetic coding and the mechanistic coding in order to address the research questions. The mechanistic coding captured entities and their activities, properties, and relationships to each other at a more detailed level than was available in the overall Russ et al. (2008) codes in order to track how students were mechanistically reasoning and to answer the first research question. These detailed codes also provided a medium for understanding change in order to answer the second research question, since they captured subtle shifts in S's mechanistic reasoning.

4.2 Pre-Interview

The pre-interview included six pre-questions and the Post Office Simulation activity which served as the first segment of Growing Certain. As discussed in the next section, the pre-questions contained a diversity of contexts which led to a variety of types of reasoning across the interview (Table 4.2). *Sample* was coded across all sections, whereas *ESD* only occurred in the sections asking specifically about ESDs such as the Post Office problem. S attended quite a bit to the data-generating process throughout, frequently questioning whether certain processes described in the problem were random even when the problem specifically stated that they were. S already showed their interest in chaining about what would happen when an additional case was added onto the sample (*Sample case is added*) in the first pre-interview question.

Table 4.2

Presence of Entity, Property, and Action Codes that Occurred in More than One Section of Interview 1 for Participant S

Entity	Property	Action	Presence in Each Interview Section						
			Hospital	Referend	Candy	Batting	PO	Casino	PO Sim
ESD			■	_	■	_	■	_	■
ESD	proportion of upper extreme		■	_	_	_	■	_	■
ESD	range		_	_	■	_	_	_	■
ESD	size		■	_	_	_	■	_	■
ESD sample			■	_	_	_	■	_	■
ESD sample	size		■	_	_	_	■	_	■
Population			_	■	_	■	■	■	■
Population	mean		_	_	_	_	■	■	_
Population	proportion		_	■	_	■	_	_	_
Sample			■	■	■	■	■	■	■
Sample	mean		_	_	_	_	_	■	■
Sample	proportion		■	_	■	■	■	_	_
Sample	proportion	changes	■	_	■	_	_	_	_
Sample	sampling		_	■	■	_	_	■	■
Sample	size		■	_	■	■	■	■	■
Sample case			■	_	■	_	_	■	■
Sample case		is added	■	_	_	_	_	_	■
Set			■	_	_	_	_	_	■
Set	proportion		■	_	_	_	_	_	■
Set	proportion	changes	■	_	_	_	_	_	■
Set	size		■	_	_	_	_	_	■
Set case			■	_	_	_	_	_	■

Note. A symbol of _ indicated that the code did not occur within that interview section, whereas ■ indicated that the interview section contained the code. PO = Post Office, Referend = Referendum, PO Sim = Post Office Simulation.

4.2.1 Pre-questions

All of the pre-questions were paper-based tasks for evaluating situations involving sample size. These questions were coded at a more general level than the Growing Certain activities, focusing on key utterances in response to each problem, and are presented in a condensed manner to provide a snapshot of S's mechanistic reasoning about sample size before starting the Growing Certain activities. Figure 4.1 shows the entities and the relationships between entities that appeared in each of the pre-questions. The Hospital problem had the most complex set of entities and relationships, with S reasoning about the causal relationship between sample cases and sample proportions, interactions with the ESD, and forming an abstract analogy of the relationship between *Set case* and *Set proportion*. The Referendum problem did not bring out much relevant mechanistic reasoning about sample size. In the Candy problem, S had a straightforward and normative chain of reasoning about the impact of a sample case on the sample mean, and how that influenced the ESD. S's reasoning about the Batting Average and Post Office problems showed a stronger causal role of the population distribution, and S returned to considering the influence of a sample case on the sample mean in the Casino problem.

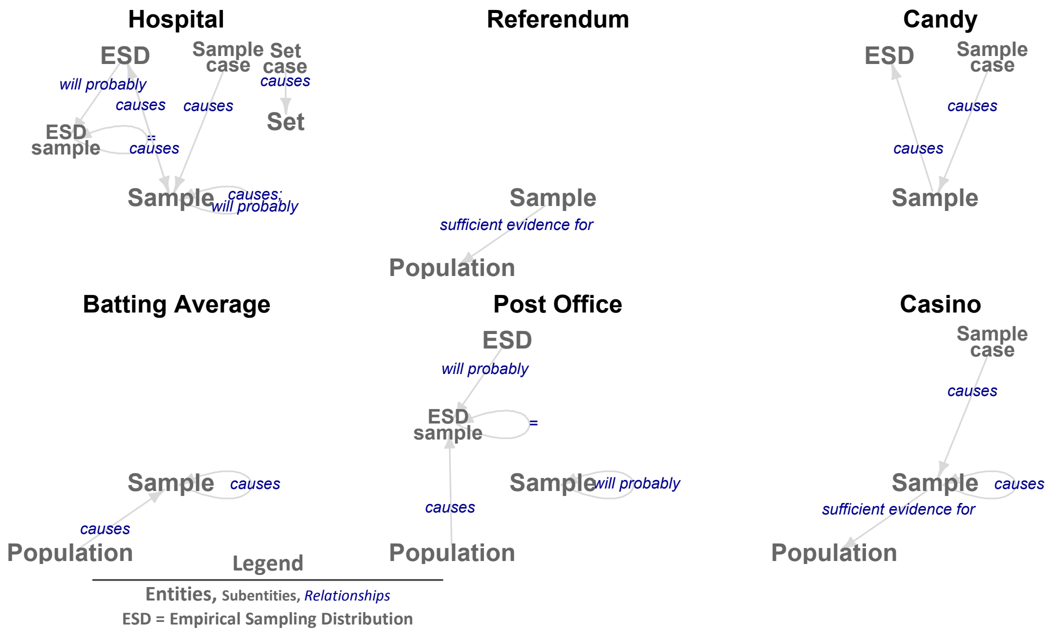
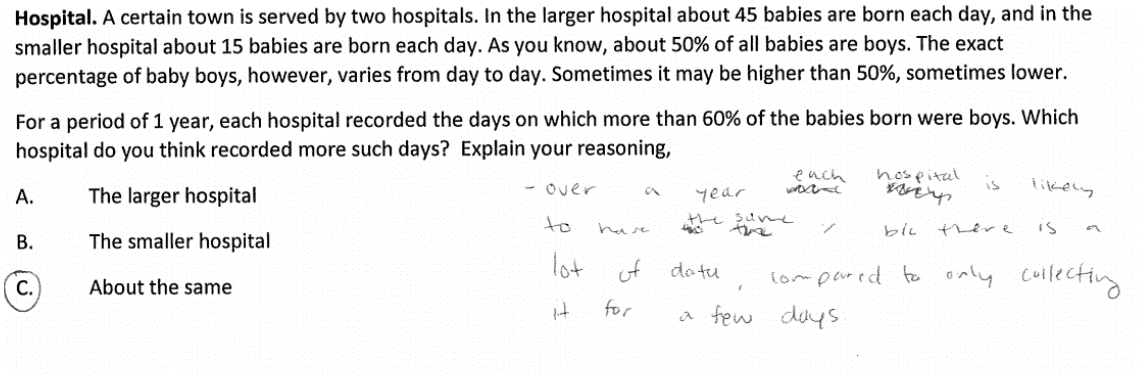


Figure 4.1. Entity and relationship summary for pre-questions for participant S.

Across all pre-questions, S displayed several instances of *chaining* about necessary connections within the mechanism, as indicated by the *causes* relationship, especially the relationship of individual cases on sample proportions/percentages which was not particularly tied to the context of random sampling. S attended to ESDs in several items that required reasoning about them (Hospital, Candy, and Post Office), and seemed to pay particular attention to the number of means in an ESD (*ESD size*), frequently conflating that with the sample size of each sample (*ESD Sample size*). Perhaps because the Referendum problem involved a setup with non-random sampling, S attended to the randomness of sampling in a couple of the later problems as well. S also seemed to pay a lot of attention in the Casino problem to how many *people* they expected to be playing the slot machine every day, which had an unclear relationship with the remainder of the mechanism of sampling variability, and a line of reasoning which would recur in the Mystery Machine activities and when S reviewed the Casino problem in the final interview.



Hospital. A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50% of all babies are boys. The exact percentage of baby boys, however, varies from day to day. Sometimes it may be higher than 50%, sometimes lower.

For a period of 1 year, each hospital recorded the days on which more than 60% of the babies born were boys. Which hospital do you think recorded more such days? Explain your reasoning,

A. The larger hospital

B. The smaller hospital

[✓ C.] About the same

- over a year ~~more~~ each ~~boys~~ hospital is likely to have ~~two~~ ~~time~~ the same / b/c there is a lot of data, compared to only collecting it for a few days.

Figure 4.2. S’s response to the Hospital problem (Kahneman & Tversky, 1972) in the pre-interview.

Many features of S’s reasoning about these problems is illustrated by how they responded to the Hospital problem, where they decided that the smaller and larger hospitals would be about the same. Even though they thought the smaller hospital was more likely to vary, the two would both be near 50% over the course of a year (Figure 4.2). S was probed on why they thought the smaller hospital was more likely to vary, leading to the following exchange:

> I: And so why is it more likely—why is the smaller hospital more likely to vary?
>
> S: Um, so if you—so if there's only 15—about 15 babies born per day, if you add one, that like s—that percentage, um, of, like, out of 15, [mh] is a—like going to increase or decrease a lot more than—so if you add 1 by 45, it's kind of like [Hm], well, if you have like a hundred people, and only 2 people don't show up, you got 80% of the class showing up, or

> something. [OK] And then if you have only, like, 10 people, and then, like, 2—or like, if you only have like 15 people and two people show up, or something like that, like, it's gonna be a much bigger difference.
>
> I: Okay. And so what—um, what answer, uh, were you choosing?
>
> S: Um, about the same.
>
> I: Okay. And so—so the kind of the second part of your explanation had to do with, um, over the course of a year, you would no longer expect to see that difference. [mh] Could you say a little bit more about that?
>
> S: Yeah, over just an extended period, I mean that's 365 days, so you get a lot more—even if you the variations, usually it'll be around—like, in total it'll probably around 50 uh, percent, just like when you flip a coin. [OK] Um, it's the same kind of percentages with that. Like, if you only do it, you know, ten times it's gonna be a little bit more varied, like, you might get, like, 70% of heads instead of more like 50. But if you do it like 100 or 300 times then you're more likely to get that—around that 50 percent. (1.237–282)

It was a bit of a surprise to hear such clear swamping reasoning, and chaining about the mechanism, emerging immediately on S’s first pre-question. S notes how a case in the smaller hospital has a bigger influence on the percentage compared to a case in the larger hospital. This seemed to be a salient aspect of percentages for S and repeatedly recurred throughout the interviews. However, S’s example involving percentages of people who show up also suggests that S’s apparently strong conceptual understanding of the swamping mechanism may be accompanied by lower arithmetic fluency, stating that 98 out of 100 is “80 percent”. S also draws an analogy a bit too quickly from the hospital problem and flipping a coin, a classic conflation of sampling and frequency distributions (Sedlmeier & Gigerenzer, 1997), where S is confusing the number of proportions drawn (365) with the sample size (10, 100, or 300 times flipping a coin), and expecting a similar evening out even though S also recognizes that the smaller hospital is likely to vary more.

S appeared to be comparing one year of births in the small hospital to one year in the big hospital, an issue that would recur in the Post Office question and which they discuss more extensively in the post-interview. Technically, the small hospital still would have a higher probability of having a percentage of boys in a year that is above 60% (≈3×10^{-50}, assuming 50% boys) as compared to the large hospital (≈3×10^{-146}), but since both probabilities are practically zero the choice "about the same" would be a reasonable answer to that question.

S's responses to the remaining pre-questions reached different conclusions, but followed similar lines of reasoning that were introduced in their response to the Hospital problem. The Referendum problem, analyzing a situation of a large biased sample from a finite population, was a bit of a departure, as S spent much of their response focused on the bias and the fact that the sample was a small proportion of the population (as did other participants). In the Candies problem, which showed the ESD directly without specifying *ESD size*, S correctly reasoned using swamping, perhaps supported by both the direct representation and the absence of a given *ESD size*. For the Batting Average problem, regarding regression to the mean for high batting averages near the beginning of a baseball season, S reasoned mostly from the fact that high batting averages are *a priori* unlikely ("the baseballs are very small, so it's very difficult to hit all of them", 1.663–664), and so a longer time playing was more likely to lead to a lower batting average. S's reasoning on the Post Office problem, which is isomorphic to the Hospital Problem but with means of a continuous variable instead of proportions, was very similar to their reasoning on the Hospital Problem, even down to the analogy of flipping a coin: "the more you do it, the more likely it's going to be, probably, um, 50-50" (1.791–793). This mention of coin-flipping and 50-50 foreshadowed their later clarity on 50-50 situations and the accompanying difficulty they had when analyzing non-50-50 situations. After some clarification on the Casino problem, meant to elicit their thinking on how to determine a sample size to determine the mean of an unknown population, S settled on 50 in order to get a "good average" (1.982), noting that there may be variability if they do a small number.

4.2.2 Post Office Simulation

Although the post office simulation was included mainly to foster comparisons with the similar simulation in Well et al. (1990), and was originally conceived as part of the pre-interview, it effectively acted as the first activity in the Growing Certain sequence because it incorporated a novel activity that involved predicting empirical sampling distributions at different sample sizes, and generating and viewing them in TinkerPlots™. Another study participant cited this activity as one of the most influential to their thinking on the post-interview, as discussed elsewhere (Brown, 2018).

S drew upon a rich array of entities and properties in this first TinkerPlots™ simulation (Figure 4.3, left), examining features of the population (*Population* and *Population Case*), samples (*Sample*, *Sample case*, and *ESD Sample*) and of abstract collections of numbers (*Set* and *Set case*). Some of this breadth of entities may be due to the fact that S actually viewed plots of the population, sample, and ESD simultaneously and was asked questions about each one of them throughout the course of the protocol. Moreover, S also showed a lot of causal reasoning across entities, particiularly between adding or subtracting the cases of a distribution and the overall mean of that distribution (Figure 4.3, right).

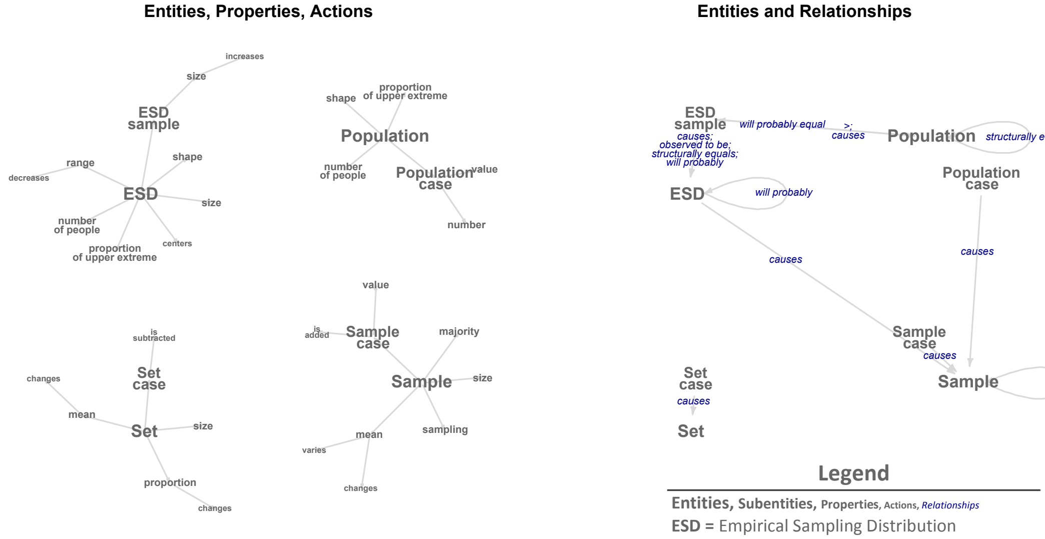


Figure 4.3. Summary of coded entities, properties and actions (left) and entities and relationships (right) for the Post Office Simulation activity for participant S.

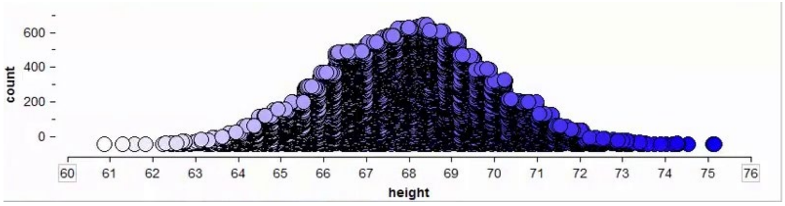


Figure 4.4. S’s TinkerPlots™ plot of the population of male heights for the Post Office Simulation.

S attended to an ambiguous property of the ESD and the Population, “number of people”, that led to non-normative reasoning throughout the activity. S brought in swamping-type reasoning quite early in this activity, even when only looking at the distribution of the population (Figure 4.4). When asked to estimate how much of the population they thought was above 72 inches, S said that “even though, like, on the screen it looks like 25 [percent], I’d probably say more like 15 or 10, just because there’s a lot more people” (1.1121–1124). When asked what they meant by this, S elaborated:

> Compared to like—if there was like a hundred people, um, that were taken—like, their data was taken, and then say, like, you could like almost individually count and also, like, one—like I said before, like one person will make a much bigger difference or impact on the, like, [Hm] say the average or what percentage it is, um, compared to if it's a thousand people it's kind of like if you lose a $1 bill and you only have ten dollars in your pocket, you're gonna b—you're gonna—it's you're gonna notice a lot easier than say if you have a thousand one dollar bills and you lose one [Hm], like it's gonna make—be much less of a difference, like, if you look at, like, say the percentages in /inaudible/. (1.1135–1152)

Apparently, because each case had a smaller impact on the total percentage, this implied that S needed to downgrade their estimate of the percentage, because of that higher total. The generality of S’s statement was quite broad, noting that in a smaller sample one person had a bigger impact on “the average or what percentage it is”—so even here, estimating the region of a dotplot or the area under a curve, this principle applied. Again, note that this principle has little to do with random sampling, and S then gave an analogy (which they return to in the next interview) that examined the percentage change of losing a dollar. S has a strong sense of mechanism here that they sometimes used quite successfully, but did not entirely find success applying this sense in a normative way.

When asked to draw what they expected the ESD with sample size 10 would be (Figure 4.5), S thought it would be “similar, if not about the same” (1.1359–1360) as the entire population, and their drawing indicated similar shape, center, and variability between the ESD and the population. S again brought in their ambiguous notion of “number of people”, noting that in the ESD with sample size ten that “I just had, like, a smaller amount of people. […] This is, like, um, the whole entire town, while this is only ten people” (1.1350–1368), even though they had clearly and correctly described earlier in the activity what one dot in each of the TinkerPlots™ plots represented, and even though they had predicted that only 1% of the ESD means would be above 72 inches as compared to 2% in the town (implying a decrease in variability that they did not carry through to their plot).

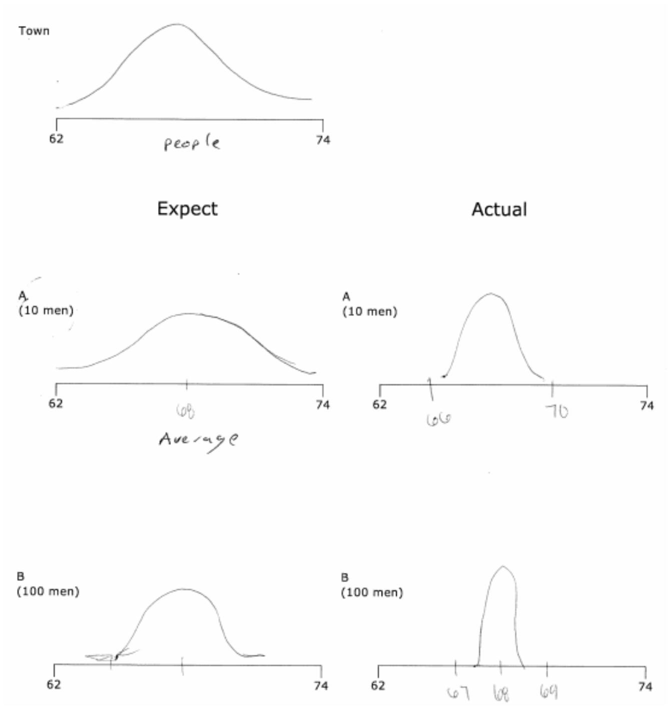


Figure 4.5. S's drawings of the heights of all males in the town (top left), expected (middle left) and observed (middle right) ESD for the means of 10 men, and expected (bottom left) and observed (bottom right) ESD for the means of 100 men. Note that S's "expected" drawing for the means of 100 men was done after they had already viewed the observed distribution in TinkerPlots™, due to interviewer error.

When they viewed the actual ESD in TinkerPlots™, they readily noticed that the range was smaller and provided an explanation based on the shape of the populations and based on sampling without replacement, similar to some of the participants in Well et al. (1990):

> Um, because most likely there's not gonna be a range of—like an average of 74, if only like maybe one or two people got 74, there's no way I can really, like, be able to go all the way over there, um, if you take like ten people. (1.1418–1424)

This was chaining about the mechanism that works based on a bell-shaped distribution and sampling without replacement, since S was essentially noting that there would not be enough people at 74 in order to fill up a sample of 10 (cf. Well et al., 1990), and thus there was no way that the mean would be 74. As seen later in the Cat Factory activity, S was able to more successfully reason mechanistically about bell-shaped distributions as opposed to other shapes of distributions. When predicting the ESD at sample size 100, S did somewhat vaguely predict that there would be fewer means over 74 than there would be for sample size 10: "maybe one percent again […] maybe slightly under one percent just because you're taking much bigger, um, amount of people and most likely it's gonna be more towards that 68" (1.1443–1448). Note that S seems to have anchored their prediction around their previous prediction for 10 men of 1 percent, not what was actually observed (0 percent). S also was for the first time predicting that a larger sample size would make the sample "more toward that 68", though whether S was clear that the ESD would actually have a smaller range around 68 is unclear, and it is also unclear whether S was more driven by the fact that the population mean is 68, or whether the population *mode*, the bulk of the values, are near 68.

Their response upon viewing the actual ESD for sample size 100 suggested that the population mode was a bigger driver of their thinking, which they chained together with swamping to explain the smaller range:

> Um, they're different because you're taking a much—a smaller sample size. So one person is gonna affect the whole entire sample a lot more, um, than, say, if you have a hundred people, you know, instead of having—like, i—it's just gonna affect it a lot less. Um, so you

> might have, like, yes you might have somebody that's 74 inches tall, but if most of the people you take are 68 [Hm], um, inches, then most likely, like, your—your mean is going to be 68, especially over 365 days. (1.1496–1507)

S repeated their familiar refrain that one person affects the sample more at smaller sizes, but now combined it with the observations that most people in the sample are likely to be near 68 because of the population shape. The sample could still draw an unlikely extreme value, but this extreme would have less effect, and combined with the population shape suggests that the mean would be closer to 68. This reasoning seems to suggest that the sample would converge to the *mode* of the distribution as sample size increases; with this particular population, S's reasoning would lead to a correct prediction because the population mean and population mode are the same. However, as S discovers later when exploring a uniform distribution in Growing Sample Means: Cat Factory 2, the sample mean will still converge even if there are multiple modes. Also, note that S again brought in the number of total means, or *ESD size*, suddenly at the end, as if this reinforced the sample size phenomenon they just observed. Since *ESD size* is arguably irrelevant here—there just need to be enough samples drawn to get a reasonable estimate of the distribution—S's use of *ESD size* as a reinforcement to *ESD Sample size* further suggests that S was conflating these two levels.

4.3 Growing Sample Proportions & Means

As the most cohesive of the 5 interviews, with a consistent focus on growing samples with known populations in TinkerPlots™, S's attention was largely on similar entities, properties, and actions throughout the interview (Table 4.3). Later in the interview, S started speaking extensively of the mean zigzagging as they saw these representations, and S's attention shifted from the proportion to the mean. When reasoning about means toward the end of the interview, S attended at several times to streaks of high or low values, which was a useful scenario to help them reason about the mechanism.

Table 4.3

Presence of Entity, Property, and Action Codes that Occurred in More than One Section of Interview 2 for Participant S

Entity	Property	Action	Presence in Each Interview Section					
			GSP Phys.	GSP TP	GSM ([illegible]	GSM ([illegible]-1	Cat Factory 1	Cat Factory 2
Population			■	■	■	■	■	■
Population	distribution		_	■	■	■	■	■
Population	mean		_	_	■	_	■	_
Population	sampling		■	_	_	_	■	_
Sample			■	■	■	■	■	■
Sample	mean		_	_	■	■	■	■
Sample	mean	centers	_	_	_	■	■	■
Sample	mean	changes	_	_	■	■	_	_
Sample	mean	stays	_	_	_	_	■	■
Sample	mean	varies	_	_	■	■	_	■
Sample	mean	zigzags	_	_	■	■	■	■
Sample	population		_	_	_	■	_	■
Sample	proportion		■	■	■	_	_	_
Sample	proportion	changes	_	■	■	_	_	_
Sample	proportion	decreases	■	■	_	_	_	_
Sample	size		■	■	■	■	■	■
Sample	size	increases	■	■	■	■	■	■
Sample	streak		_	_	_	■	_	■
Sample case			■	■	■	■	■	■

Note. A symbol of _ indicated that the code did not occur within that interview section, whereas ■ indicated that the interview section contained the code. GSP Phys. = Growing Sample Proportions: Physical Simulation; GSP TP = Growing Sample: Proportions TinkerPlots™; GSM = Growing Sample Means

4.3.1 Growing Sample Proportions: Physical Simulation

S again already displayed rich mechanistic reasoning about relationships between individual cases and the movement of the sample proportion as sample size increases when viewing a sample grow from 1 to 10, perhaps due to their strong intuitions about 50-50 situations. S started the segment attending more to aspects such as *Sample sequence* and the how the *Population sampling* was random (Figure 4.6, left), but by the end of the segment was again describing in detail how a single *case* of a *Set* or a *Sample* caused changes in the *proportion* when it was added or subtracted, and that the impact would be different at different sample sizes (Figure 4.6, right). The *Set* entity was again within the context of S's wallet analogy, which S elaborated on extensively in this interview. Also visible on Figure 4.6 is S's first mentioning of the *Sample number of slots*, where S noted that not all slots—possible values of the proportion—were possible and hinted at understanding of what those slots would be at a sample size of 10. S's attention to slots would become a major theme of their thinking in the later interviews.

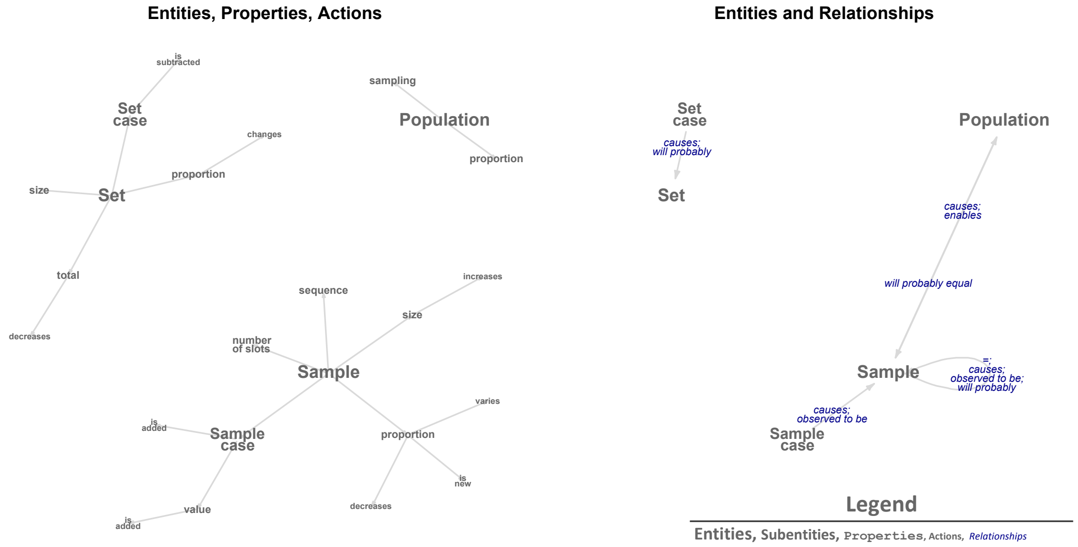


Figure 4.6. Summary of coded entities, properties and actions (left) and entities and relationships (right) for the Growing Sample Proportions: Physical Simulation activity for participant S.

After being introduced to the situation involving picking one of two blocks from a box 10 times, and tracking what was drawn and the total blue and percentage blue so far, S was asked what they would expect the table to look like. Their initial response indicated a lot of uncertainty about the form of the table:

> Um, probably, like, just random. I don't really know—like, if it was—it'll be—it—it like might be close to 50, but also, like, it could vary just because [OK], like, it's random, so. (2.33–37)

They here started displaying some idea that there may be a tendency to be near to 50%, but without much commitment to what could happen on the table. They both said that the table would *look* "random", and that it could vary because the process is "random". Upon further probing, they expanded on their use of the word "random":

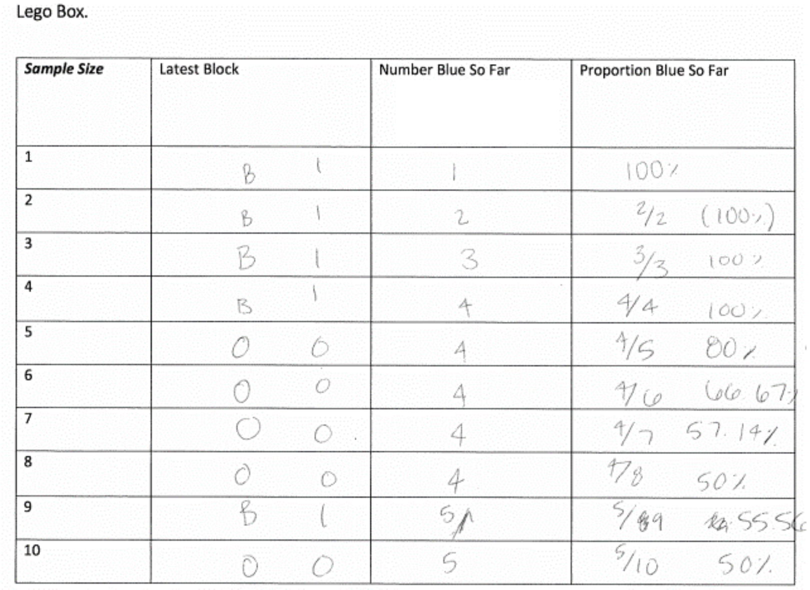
> If we did more trials than 100—or like than ten [Hm], but I'd say, like, maybe even, maybe more, maybe less. [OK] So, [OK] I can't really predict if it's gonna be random. But if it was a perfect scenario, then it would be like 50 per—like, 50-50 [OK] but. (2.43–49)

Unprompted so far on this particular interview, though perhaps cued by the focus on sample size in the first interview, S identified the property of sample size. Here, their use of the word "random" seemed to denote "unexpected", and S contrasted this with the "perfect scenario" where the proportion blue is exactly equal to 50 percent.

However, the dialogue took a turn when S was asked what will happen as sample size increases—the first time in this interview where the code *Sample size increase* occurred, which perhaps shifted attention to a "growing a sample" perspective. Their answer focused on the fact that the first block drawn was a blue, and so the sample proportion was already at the upper extreme. Therefore the proportion could only decrease, and they linked this with the batting average problem, which also focused on a decrease in an extreme proportion as the sample size increased. Already, before any physical or simulated sampling unfolded, S's attention to the mechanism seemed to depend on the prompts that they were given. After an initially vague response that mentioned 50-50, and then seemed to distinguish between whether an outcome was "random" or "perfect", attention to *Sample size increase* led S to make a connection with the batting average problem and to articulate the principle of regression to the mean:

> Like the—the baseball example last time, how like [OK] as the season goes on and they hit more, their, like, um, batting average usually decreases. So I would say that the proportion blue so far would decrease over the sample—or, like, however many times you do it. (2.64–70)

Because the first block happened to be blue, this may have helped S draw this analogy to the batting average scenario, which also involved a high extreme statistic decreasing as sample size increased.



Lego Box.

Sample Size	Latest Block	Number Blue So Far	Proportion Blue So Far
1	B 1	1	100%
2	B 1	2	2/2 (100%)
3	B 1	3	3/3 100%
4	B 1	4	4/4 100%
5	O 0	4	4/5 80%
6	O 0	4	4/6 66.67
7	O 0	4	4/7 57.14%
8	O 0	4	4/8 50%
9	B 1	5	5/~~8~~9 ~~ka~~ 55.56
10	O 0	5	5/10 50%

Figure 4.7. S’s outcomes of drawing 10 blocks from a box containing 1 orange block and one black block.

After actually observing 10 blocks, however, the fact that a running total of proportions led S to a new observation: not only was there in fact a decrease, but the amount of the decrease was smaller after the initial decrease (the decrease was decreasing). By chance, the first four blocks drawn were blue (Figure 4.7), so S’s attention remained on the decrease, though there is not evidence that S was viewing decreases as substantially different from increases or moving back and forth: They noted that it “kept going down, like, less" (1.143–144) which was relatively descriptive for this sample. When the interview probed for why S thought this, they directly attributed the change to sample size, and explained using an abstract analogy:

> Because you have a bigger, like, sample size. So it's kind of, like, if you have like a wallet full of like $1000—like a thousand, like, one dollar bills and you take one out [mh], you're gonna probably forget about it [Hm] or, like, it really won't make a difference in the long run, while, like, if you only have $10 and have like—of $1 bills, and you have—take one away, like, you're gonna have a lot—like, you're not gonna be able to pay for a lot less. (2.151–162)

Note that this wallet example has nothing to do with any sampling process, and the properties that were noted here do not map clearly on to means of random samples. The reasoning subtly changed as their explanation continued. At first, they noted that someone with $1000 who loses $1 will not notice, whereas someone with $10 who loses $1 will, without explaining quite why this would be the case. Initially it seems as if S described something like the size effect from numerical cognition research (Moyer & Landauer, 1967), that people have more trouble perceptually differentiating 999 from 1000 than they do differentiating 9 from 10. There was no sense of how the $1 difference changes between a *small Set size* and a *large Set size*. But then S invoked proportional reasoning in their analogy:

> But, just in general, like, one out of like—if you're like 999 dollars out of a thousand, like, the percent it's not really gonna change that much, [mh] just because of, like, how big of it—how big it is, while, like, having 9 out of 10, that's, like, a much larger difference. (2.163–170)

Here S appeared to be making the observation that removing one dollar out of a thousand has a smaller percentage change than removing one dollar out of 10—a variant of swamping reasoning. A couple things are confusing about this analogy, including that it is not clear what "percentage" S is talking about and how that maps on to samples. Perhaps S is talking about the percentage change in number of dollars—the percentage change in the *Set total*.

Although S had previously applied similar reasoning about the difference in impact in one case between large and small samples, S now directly applied this reasoning to explaining the decreasing amount of change in proportion as the sample size grows. S's attention to the decreasing amount of change may have been enabled by the representation of the table, which foregrounded the phenomenon of decreasing change as opposed to a static contrast between large and small samples. This appeared to enable S to make a more precise statement of the mechanism regarding the effect a single case has on the proportion at different sample sizes, now that S had observed that phenomenon directly.

When asked about what would happen as the sample size increases, however, S no longer attended to swamping mechanisms. Instead, perhaps inspired by the fact that the sample was already at 50% (Figure 4.7), S hypothesized that "it would probably stay around 50 […] because—um, you kind of have, like, a 50-50 chance" (2.176–179). S here made the link to the setup conditions more explicit, and was perhaps cued to attend to this because the sample had already ended at 50%. S noted that they would still expect the proportion to "go up and down" (2.179) but that "if you did, like, say 500—it 500 times, like, it'd probably stay around 50" (2.181–183). S mentions both the impact of a given case on the proportion, and then predicts that the proportion will stay around 50 at a large sample size, but they do not make an explicit connection between these two assertions. This fuzzier mechanistic reasoning about the expected proportion blue of a much larger sample, after a very specific and well-articulated chaining about the impact of the case on the proportion in S's wallet analogy, may be due to the fact that S's swamping mechanism at this point was not well-connected to any features of the random sampling and was simply a property of sets. While S clearly realized that there was something important about 50% and attended to the setup conditions here, they were not yet able to use that information to reason deeply about the mechanism, running up against the limits of swamping for explaining the Empirical Law of Large Numbers.

After attending to the changes in the proportion as the sample size grew from one to ten, S then did have a more specific prediction of what might happen if S were to grow another sample from 1 to 10, stating that "at the very end it'll probably be maybe arou— like, give or take, like, say, like, 10% for 50, just because of, like, how many, like, slots there are, so" (2.196–200). It is not clear what led to this statement. S perhaps was noting that their actual final proportion was 50% which matched the setup conditions. However, this prediction was immediately after statements about how adding on individual values would increase or decrease the proportion, and so there appeared to be more than just attending to the setup conditions. A possible explanation here is that the "slots" are the possible values that the proportion can take on at every sample size, and therefore a "give or take" of 10% makes sense for sample size 10, because both 40% and 60% are possible values of the proportion. The "slots" may also correspond to the "decreasing less" that S already noted. Therefore, S may be coordinating both swamping—which does map dynamics as sample size increase that S observed—and an awareness of the setup conditions. “Slots” could be considered a middle ground between swamping and heaping: The attention to possible outcomes is related to heaping, while the decreasing distance between the slots is essentially the phenomenon of swamping. S does not make these connections explicitly here, but does expand on their reasoning about slots in the later interviews.

Throughout this first activity, S also seemed to attend to how the individual outcomes chunked, perhaps highlighted by the table representation (Figure 4.7), the fact that they had been asked about whether they expected patterns in that regard, and the fact that this sample actually started with four blue blocks and so the proportion stayed at 100% during that time. When S was explaining about the slots, they started by giving an example of what might happen if it started “all orange” or if there was a “large portion” of one color (2.192–195). Their last thought on what would happen if a new sample were drawn was that they wouldn’t necessarily expect the outcomes to come in “two chunks” (2.209) as this one did. It is not clear how this attending to chunking as a part of understanding random behavior may have affected other parts of reasoning about the mechanism. As the interviews progressed, however, S did tend to use streaks of similar outcomes frequently in their reasoning, perhaps because these streaks had clear and visible effects on the mean in the sample size plot.

4.3.2 Growing Sample Proportions: TinkerPlots™

After viewing a single simulation with sample size 10, with some highlighting of the changes in the proportion blue, S's predictions had already become more specific and they expressed some connections to swamping. At this point, S transitioned to reasoning with the sample size plot and viewing simulations of larger sample sizes in TinkerPlots™.

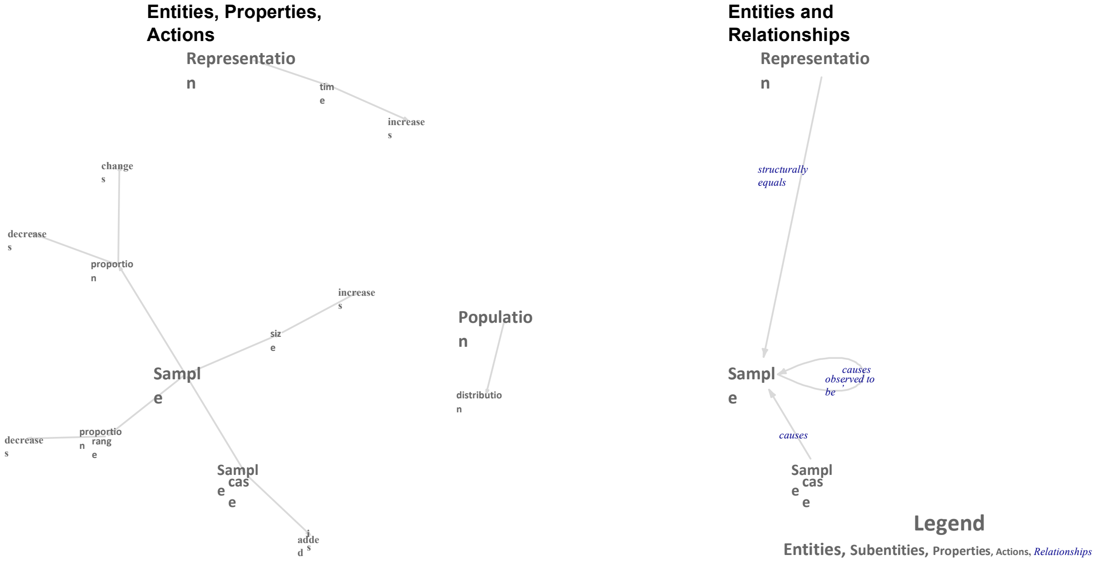


Figure 4.8. Summary of coded entities, properties and actions (left) and entities and relationships (right) for the Growing Sample Proportions: TinkerPlots™ activity for participant S.

Perhaps since this was S's first exposure to the sample size plot, S spent some of the time focused on how the increases of time on the *Representation* as they continued growing the sample corresponded to the increase in the sample size (Figure 4.8). S's attention to properties of the *Sample* corresponded closely to what the sample size plot made visible: additional *Sample cases* being added onto the sample and causing the *proportion changing* and the *proportion range descreasing*. Although S vaguely mentioned the distribution of population values, they did not explicitly relate the population to the sample in this segment.

The initial exposure of the sample size plot—the line plot with mean or proportion on the x-axis and sample size on the y-axis—was meant to introduce the basic mappings of the representation without revealing the distinctive trend of decreasing variability as the sample size increases. However, when S was asked to produce what they thought the sample size plot of the orange and blue block would look like, they struggled and initially drew a bell curve. I provided some more guidance, adding more made-up numbers to the TinkerPlots sampler to show that the plot should start at the bottom and go up as sample size increases. At this point S noted that the plot should be "like a zig-zaggy…" (2.282), a theme that they would frequently return to while reasoning about the shape of the plot as it related to sample size. Over the original bell curve drawing, S then drew a zigzagging line with perhaps some slight change in variability (Figure 4.9, left), though much less than the typical simulated line.

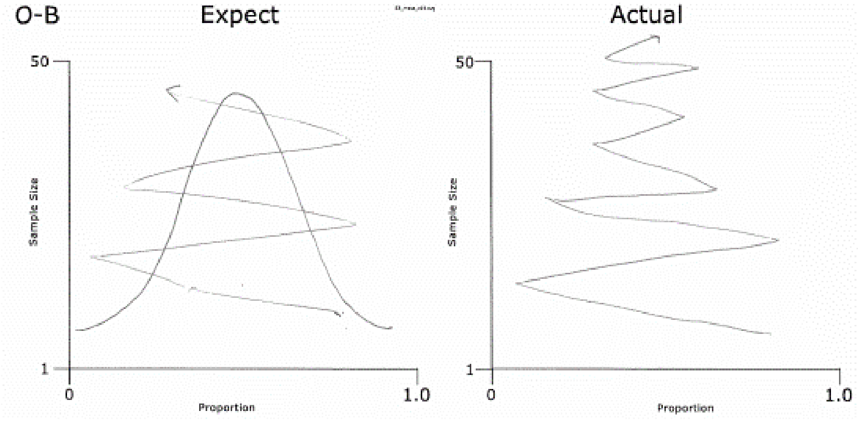
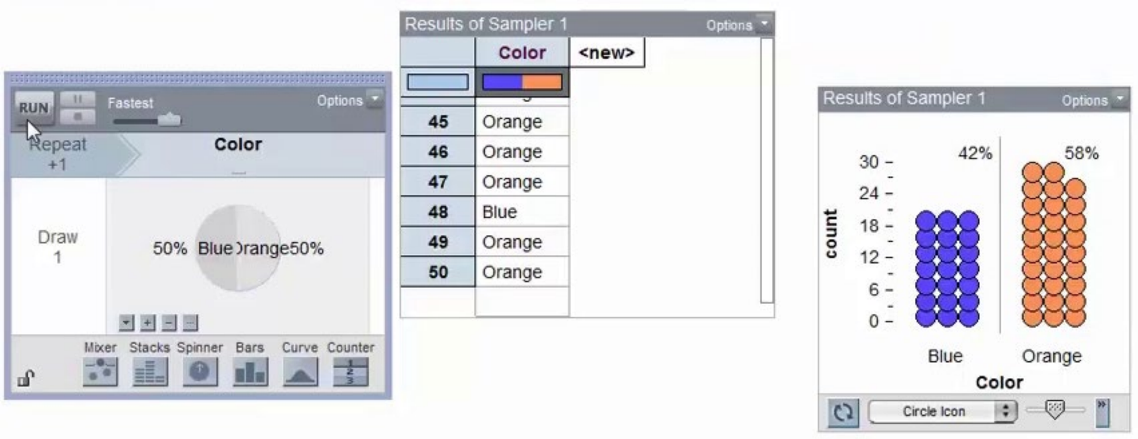


Figure 4.9. Predicted (left) and observed (right) sample size plots for Growing Sample Proportions, plotting the proportion blue drawn as the sample size grows up the y-axis. S initially drew a bell curve in the predicted panel before drawing a line plot over it.

Perhaps because of the newness of this representation, and their initial confusion of how to draw the graph, S's initial description of the graph simply described how moving up on the graph corresponds to growing the sample size. S vaguely described how "it'll kind of bounce between the two sides" (2.316–317), echoing their previous comment about the zig-zag. They commented that "for a proportion of 0.5, it'll probably kind of stay around 0.5, maybe, and then just kind of go up" (2.321–324). The phrase "for a proportion of 0.5" may indicate S's recognition that the chances are 50-50 for orange or blue, and their trying to work that in to this situation somehow, but the connections between the zigzagging, the sample size increasing, and the settling down to a long-run proportion of 0.5 are not very clear or integrated.



	Color	<new>
45	Orange	
46	Orange	
47	Orange	
48	Blue	
49	Orange	
50	Orange	

Figure 4.10. S's TinkerPlots setup and results for drawing 50 orange or blue blocks, with equal probability, one at a time.

S was easily able to set up the sampler for the block situation in TinkerPlots™, which is a similar 50-50 situation to the coin flip simulation in CATALST's Introduction to Monte Carlo Simulation activity (Zieffler & Catalysts for Change, 2017, pp. 24–29); however, setting up the sampler to incrementally add on to the sample was clearly new to them. Although S drew a sample that ended with only 42% blue blocks (Figure 4.10), this did not appear to perturb them. Based on just looking at the percentage as the sample size grew, they again commented about the proportion "kind of going back and forth" (2.438), but this time noting the smaller size of the changes:

> Once we got like more times it started just decreasing, like, less. So instead of having, like, say, like, a 10% decrease or increase [Hm], like, it went more of, like, 2% or 5%, or something like that. I wasn't looking at the exact percentages, but. (2.444–449)

Again, they appeared to recognize that the proportion should end up near 50%, even though it in fact was at 42%:

> then it kind of started going more towards, like—instead of, like, say, like, 80%, it was more centered around, like, 50/50. (2.450-453)

Structurally, their response mirrors what they predicted—first, commenting about the zigzagging, and then about the eventual central tendency. This same sequence of reasoning, first to the impact of individual cases and then to the central tendency, was also observed in their reasoning about the table of 10 blocks, above. One explanation for this repeated structure is that these representations make the moment-to-moment changes in the proportion of primary salience, but S's strong intuition that the proportion should be near 0.5 also is expressed after they have made the initial comment about the decreasing size of the variability.

After actually drawing the sample size plot of what they saw, based on their memory of how the percentages moved, S commented that "it's a bigger range in, like—w—in, like, the zigzagging, and then it kind of, like, gets smaller as you go up" (2.470–474). Both the terms "zigzagging" and "range" did not appear in their explanation right after viewing the simulation but before the sample size plot. The decreasing zigzags are a distinctive signature of swamping, and in fact S went on to correctly articulate the differential effect of each case on the proportion in small and large samples. S's reference to range, however, echoes their language when describing what they saw during the Post Office Simulation, noting the range of the ESD decrease as the sample size increased. This provides some hint that the sample size plot may have a use as a bridging tool, or a complementary tool, for seeing how ESDs decrease in range, since there is some sense of range mapped in the sample size plot.

4.3.3 Growing Sample Means: 0-1

The next section shifted the focus from proportions to means of a 0-1 variable, scoring the orange block as 0 and the blue block as 1. Although the mean was the same as the proportion blue, this equality was not mentioned or emphasized. Additionally, the sample size was grown all the way to 200 and S actually saw the sample size in TinkerPlots™ as the sample size grew. Although S never stated that the mean was the same as the proportion, many of the relationships and codes observed during this segment (Figure 4.11) paralleled those in S's reasoning about the Growing Sample Proportions: TinkerPlots™ segment (Figure 4.8). Note in Figure 4.11, left, the presence of *Sample mean zigzags*, mirroring references to *Sample proportion zigzags* earlier. In fact, S once slipped into discussing the *Sample proportion* when describing the sample size plot. S started to connect the results in the sample a bit more specifically to the population, noting a causal role of the Population setup on the sample mean for a large sample (Figure 4.11, right).

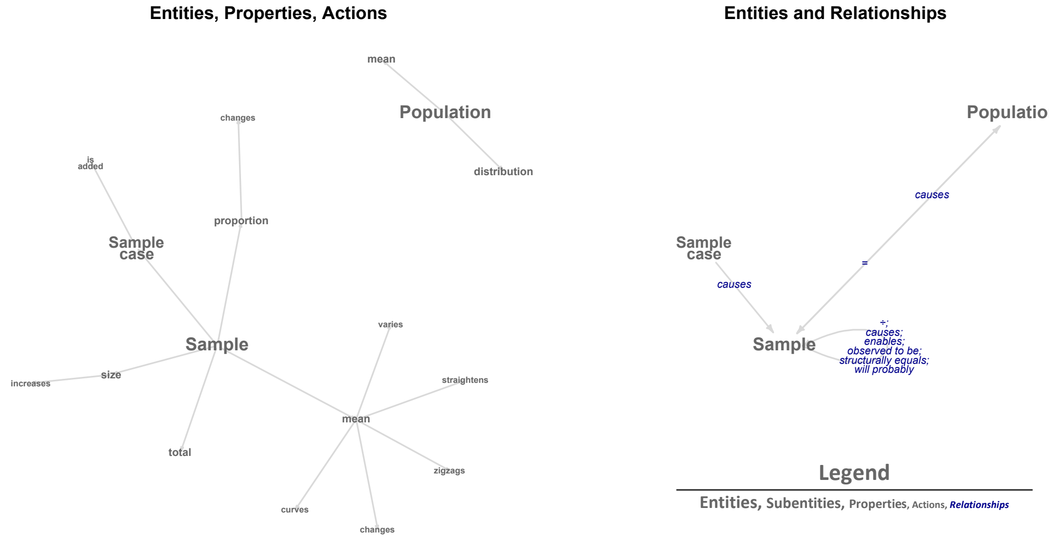

Entities, Properties, Actions
mean
Population
distribution
changes
is added
proportion
Sample case
Sample
size
increases
total
varies
straightens
mean
zigzags
curves
changes
Entities and Relationships
Population
causes
=
Sample case
causes
Sample
÷; causes; enables; observed to be; structurally equals; will probably
Legend
Entities, Subentities, Properties, Actions, Relationships

Figure 4.11. Summary of coded entities, properties and actions (left) and entities and relationships (right) for the Growing Sample Means: 0-1 activity for participant S.

Note also the presence of the activity *Sample mean curves* (Figure 4.11, left). Interestingly, S's expected sample size plot for the 0-1 mean situation was a smooth curve from 1 and bending towards 0.5 (Figure 4.12). It is unclear why S's focus was now away from the "zigzag" that was so salient previously—including on S's drawings of the expected and actual proportions for the same situation (Figure 4.9), which was actually on the same piece of paper that S was drawing for the mean. One possible explanation is that since the interviewer lengthily explained how to calculate the mean at each sample size, S focused more on structural features of the problem and only attended to the long-run expectation of 0.5 rather than the path it would take to get there. It is also possible that something about the statistic of interest being presented as a mean may have altered S's view of the situation, even though they still recognized the long-run expectation of 0.5.

Either way, it seems that S's awareness of the phenomenon of the sample statistic zigzagging as the sample size grows was relatively fragile at this point.

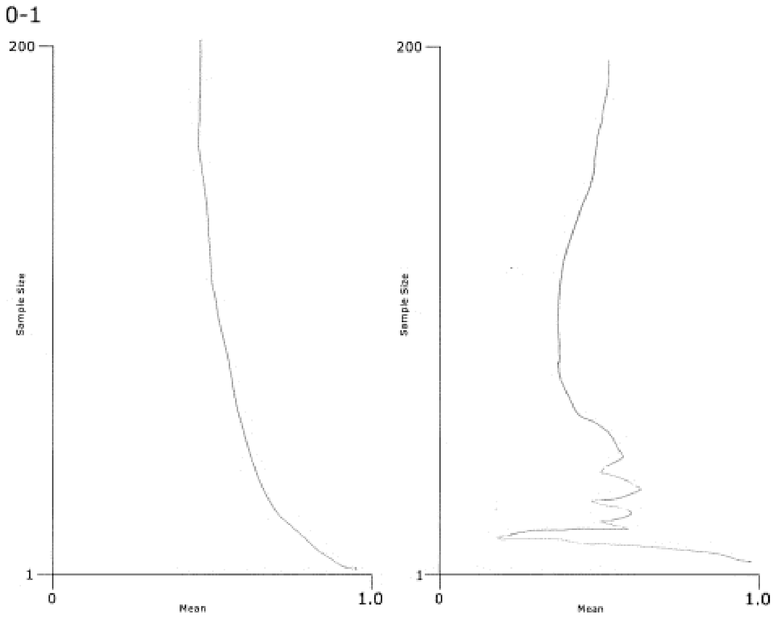
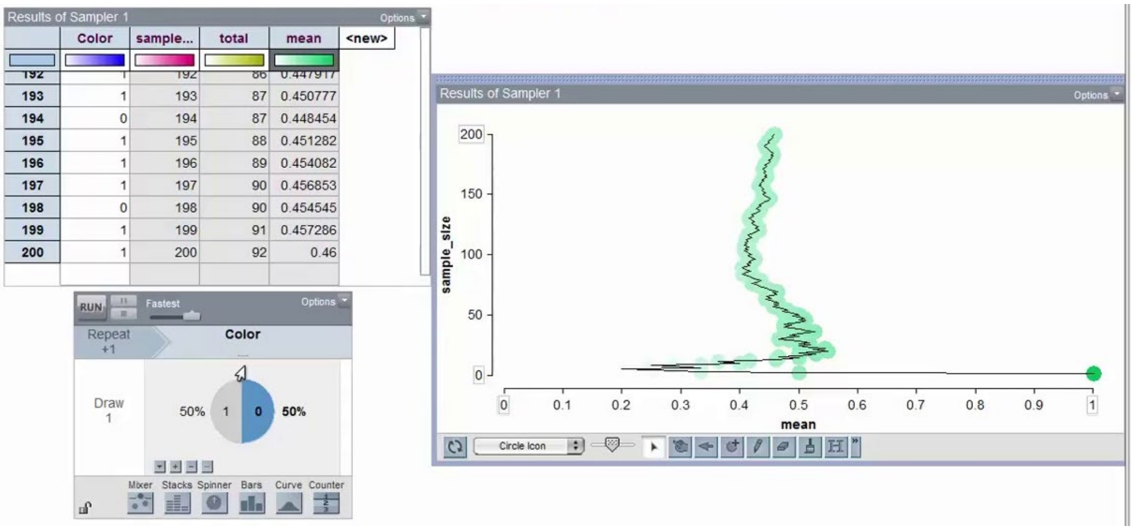


Figure 4.12. S's expected (left) and actual (right) sample size plots for the Growing Sample Means: 0-1.

S attended to structural relationships as a part of setting up TinkerPlots™ to show the sample size plot, which required some formula calculation to create the running total and the running mean for each sample size. After some prompting, S recognized that the mean for each sample size would simply be the total blue divided by the sample size and created that formula in TinkerPlots™, and then created the sample size plot (Figure 4.13).



Results of Sampler 1

	Color	sample...	total	mean	<new>
192	1	192	86	0.447917	
193	1	193	87	0.450777	
194	0	194	87	0.448454	
195	1	195	88	0.451282	
196	1	196	89	0.454082	
197	1	197	90	0.456853	
198	0	198	90	0.454545	
199	1	199	91	0.457286	
200	1	200	92	0.46	

Figure 4.13. S's sample size plot in TinkerPlots™ for Growing Sample Means: 0-1. The sampler, lower left, shows 0 and 1 each with a 50% probability of being chosen, and set up to add one value incrementally on to the sample; the sample, upper left, shows the value drawn, the sample size, the running total of 1s drawn so far, and the running mean so far; the sample size plot, right, shows the mean (x-axis) against the sample size (y-axis).

When commenting on the actual results, S did mention that "it's kind of gonna—maybe to—maybe shoot over at one point, like, /laughs/" (2.734–735), an apparent reference to the large decrease from 1.0 to 0.2 from sample size 1 to 5, and a contrast to the smooth curve that S had drawn as an expectation. S quickly shifted focus, again, to the long-run tendency, but now was more specific:

> And then it's gonna stay around, um, I'd say, like, it'll kind of be like around 50 after about, like, it seems like at least on this graph, like, it's gonna be like around like 25 or 30 [OK], and then it's kinda gonna just kind of stay in the middle. [OK] So. (2.736 –740)

Here, for the first time, S identified that the long-run tendency for the mean was first seen at a specific sample size. This observation is important for inference, since a sufficient sample size is needed to have enough power for precise estimation. S was not specific about how much it would "stay in the middle", and in fact quite a bit of variability was still observable in the sample they drew (Figure 4.13).

Another notable feature of S's reasoning at this point is that they were already using their results to predict what they would expect to happen in general, using language like "it's gonna be" (2.733). Note that when describing how there was a sample size at which the mean stays in the middle, S said "it seems like, *at least on this graph*, like it's gonna be around like 25 or 30" (2.738–740, emphasis added), with the implication that this graph was a particular instance of a more general phenomenon. In prior moments of the interview, before being prompted to predict, S simply described what they saw as a result of the 10 physical blocks being drawn or the TinkerPlots™ simulation of drawing the orange and blue blocks. There are several possible explanations for this change. One is that all the tasks repeatedly asked S to predict—before doing the simulation, and afterwards what they would expect if they did the simulation again—and S may have been catching on that predicting what would happen was a major goal for these tasks. Another possibility is that this was the first task where the simulated representation in TinkerPlots™ was the same as the representation S used to predict—S was for the first time seeing the sample size plot directly produced, when the sample size plot was previously used as a predictive tool or as a way of summarizing what they thought they saw. These data cannot distinguish between these possible explanations (or a combination of both).

S did mention that the result was different than their expectation and that they had not originally expected the mean to zigzag, consistent with the fact that they drew a smooth curve. Their explanation contains some clues as to what they recalled about their original thinking:

> I didn't really think about, like, how it's, like, definitely gonna either start at one or zero, can't really start, like, in the middle. [mh] So, just how it's gonna start. And then probably zigzag in the beginning [mh], but it makes sense. [OK] I just didn't, like, think about the— I did more of, like, well it's about gonna be like this, and not, like, really zig zag. So. (2.755–764)

Their expectation, in fact, did actually start at 1, not "in the middle", and back when they were making their expectation, S asked "does it really matter what side I start from?" (2.549–550). However, S's attention to the structural relationships, here the necessary fact that the mean at sample size 1 must be 0 or 1, may have represented a fuzzy realization that many of the values at the beginning of their expected sample size plot (Figure 4.12) were simply not possible values. This reasoning about possible values was consistent with S's previous comment in Growing Sample Proportions: Physical Simulation that they would expect the sample mean to be within ±10 percentage points of 50% because of "how many, like, slots there are" (2.199–200). This reasoning about possible values appeared to lead to the zigzagging that S had failed to predict this time, even though they had predicted the zigzagging in Growing Sample Proportions: TinkerPlots™ Simulation.

When asked to predict what would happen in a new sample, S again noted the change in the zigzagging as sample size increased, with more precision than they expressed during the previous task: "it's probably gonna still, like, be a lot more zigzaggy at the bottom that it is, like, at the top, where it’s kind of gonna just taper off into a straighter line” (2.771–775). Interestingly, S did not note the central tendency at this time, perhaps because the recent conversation was more focused on the zigzagging. S’s statement here is the clearest statement yet of the phenomenon of the relation between sample size and the movement of the mean, with lots of zigzagging at the bottom and then gradually becoming a “straighter line”, probably supported both by actually seeing the representation for the first time in TinkerPlots™ as well as the repeated questioning about this topic which supported the clearer articulation.

When asked why they expected more zigzagging at the bottom, S was also able at this point to describe chains of reasoning for the mechanism by which sample size affected sampling variability:

> So each time you put, like—have one more trial, um, it's gonna have, like, one less—or I mean it's gonna have a bigger impact on the total, like, percentage or, like, say on—in this case, like the mean. Um, it's just gonna be able to vary a lot more, while if you have like 200, then, like, adding one blue or one orange is gonna make less of a difference in, like, say the percentages of each one, so. (2.782–791)

This was not new reasoning for S, and still was not strongly connected to anything related to random sampling. S’s reasoning appeared to be more about percentages than about means: S first mentioned the “total, like, percentage”, only parenthetically recognizing that it would be the mean “in this case”. Again, S concluded by discussing “the percentages of each one”. Although this reasoning is of course correct in this case, it is not clear whether S would have been able to volunteer this kind of swamping information in a situation where there were not percentages involved. A notable difference in S’s reasoning here, however, is S added the concept of “gonna be able to vary a lot more”, coded in Figure 4.11 as the *enables* relationship. Not only does sample size play a causal role, but it also plays a role in *limiting* how much a new case can change the overall mean. This hint at possibilities, again echoing S’s previous mention of “slots”, is another foreshadowing of the kind of reasoning necessary for understanding heaping.

4.3.4 Growing Sample Means: 0-1-1

So far in this interview, S had only been working with equiprobable variables, which they seemed relatively familiar with given their frequent comments about expecting it to be “even” and “50–50”, all the way back to their discussion in the Hospital problem in the first interview to how “it’s like when you flip a coin” (1.205). When another blue block was added into the box, S struggled to articulate what would happen to the central tendency when a one was twice as likely as a zero, and never articulated the population mean (Figure 4.14, left). S did predict that the mean would be centered higher but never gave a specific value, and struggled to set up the TinkerPlots™ sampler. S generally attended to similar entities and properties as in the 0-1 situation (Figure 4.11), but no longer made reference to proportions, perhaps because the situation was no longer isomorphic to coin flips and S lost access to that resource for their reasoning. Many of the codes for this section revolve around the difference in central tendency, but S did also note that the 0-1-1 situation had less variability at the bottom than did the 0-1 situation and engaged in complex chaining about the mechanism for why this would be true, expanding their reasoning about *streaks* of values within a sample.

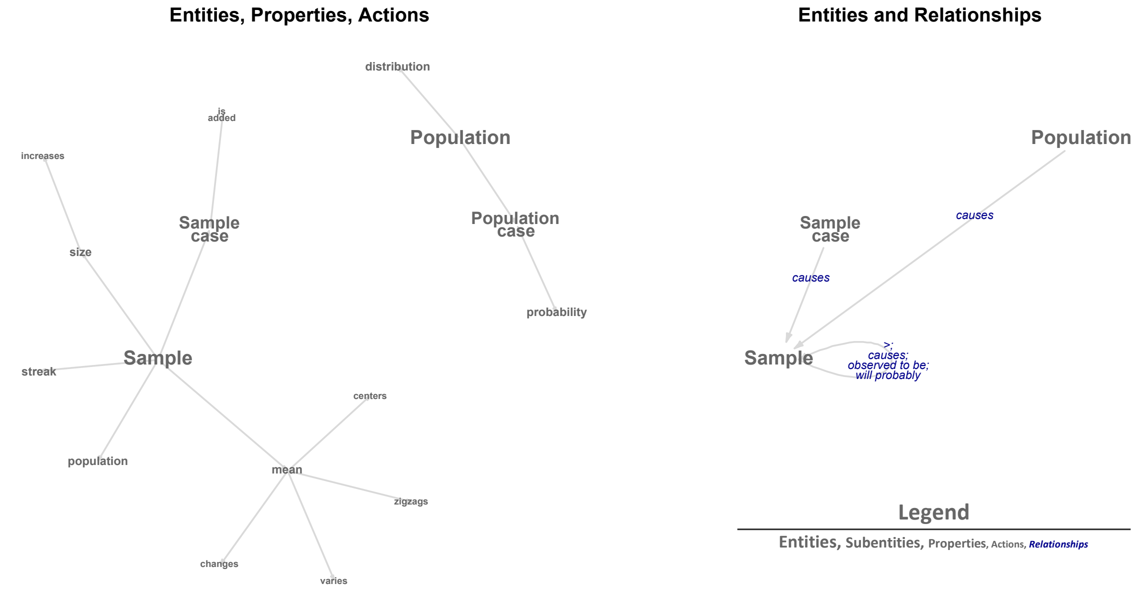


Figure 4.14. Summary of coded entities, properties and actions (left) and entities and relationships (right) for the Growing Sample Means: 0-1-1 activity for participant S.

This unclearness about the center was present upon being introduced to the change in the population distribution. S's attention was immediately on the change of where they expected the graph to be centered, predicting that "it's gonna be more close to, like one, say, than in the s— in—in—in the middle, so" (2.807–808), and then again reiterating that it would be "centered on, like, the right side of the graph" (2.819–820). This unclearness about the central tendency may have stemmed from the same source as their unclearness about how to set up the model in TinkerPlots™:

I: Okay. Um, so how can you change the model on the TinkerPlots™?

> S: Um, should I just—like, could I put, like, 75 and 25%, is that the right percentage, =or—or—=
>
> I: =So there's= two blue blocks and one orange block.
>
> S: I guess we could do the proportion. /*Changes TinkerPlots™ sampler to display proportions*/ Would that be betterrrr? Or… my brain is a little fuzzy right now. /*Changes TinkerPlots™ sampler back to percentages*/
>
> I: That's okay. That's okay. So, um—
>
> S: It'd be like—s—/*starts editing percentages in sampler*/
>
> I: What's your chance here of drawing an orange block?
>
> S: /*hesitantly*/ 33%.
>
> I: Okay.
>
> S: So then. [OK] All right. (2.826–854)

I was caught off guard by S's difficulty in defining the TinkerPlots™ model in contrast to the ease with which they had created the sampler in the 0-1 situation, and I accidentally slipped into a teaching mode, albeit a Socratic one. S knew that the percentage of ones should be more than 50–50 in their TinkerPlots™ Spinner, but was not certain about what the percentages should be. When I gave a perhaps overly evaluative reminder of the setup of the situation, S wondered whether they should change the percentages in TinkerPlots™ to proportions. Only when I directly asked about the expected population frequency for an orange block, then they responded correctly and used that as the percentage of 0s. It is possible that S may have had difficulty setting up the simulation because they decided to use a Spinner device in TinkerPlots™ (Figure 4.13), which represented the probability as a circle and thus required S to calculate the probabilities in order to represent the process. An alternative and probably easier solution would have been to use a Mixer, which would have allowed S to simply put a 0 and two 1s in the Sampler and which was more directly analogous to the physical simulation of blocks sampled from a box with replacement. However, S displayed similar confusions about the 0-1-1 process when interacting with a pre-built TinkerPlots™ model that used a Mixer in later interviews (Figure 4.37).

Although S mostly discussed the change in center, they did predict that “the zigzag will probably be, like, the same. Like, it’s still gonna vary a lot in—at the very bottom” (2.820–823). This was reflected in their expected graph (Figure 4.15). S continued their uncertainty about the eventual center while drawing this graph, indicating that ’it'll probably be like... sh.... or, yeah, close enough.’It's closer to one. Maybe—not like exactly one, but /*laughs*/ yeah” (2.881–883), and indicated they felt they drew it a bit too far to the right.

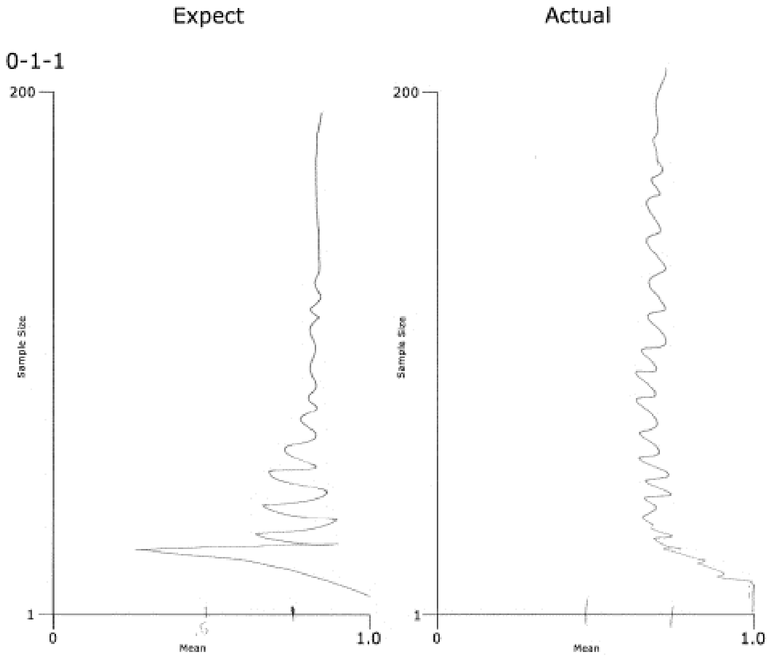
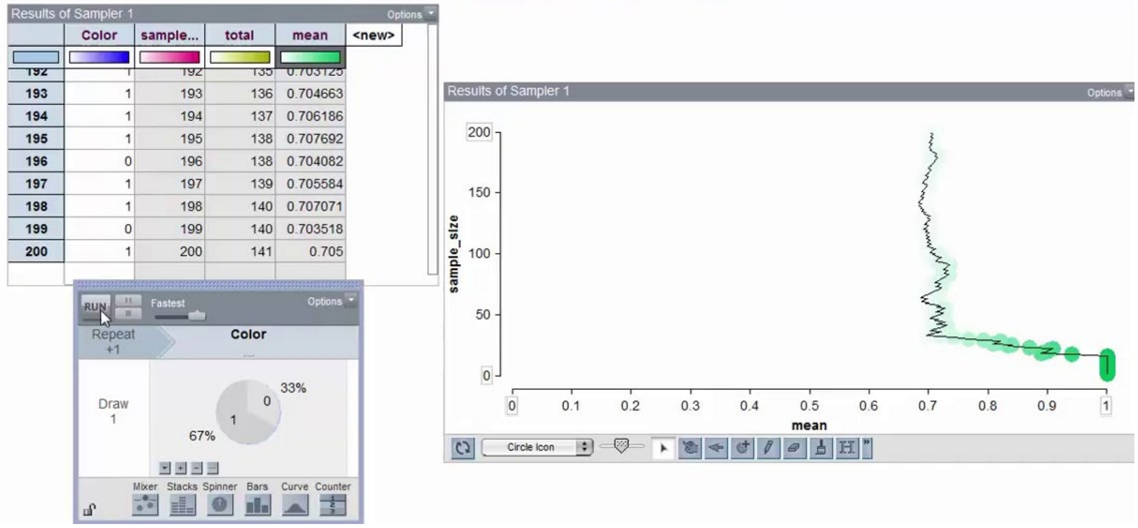


Figure 4.15. S's predicted (left) and observed (right) sample size plots for Growing Sample Means: 0-1-1.



Results of Sampler 1

	Color	sample...	total	mean	<new>
192	1	192	135	0.703125	
193	1	193	136	0.704663	
194	1	194	137	0.706186	
195	1	195	138	0.707692	
196	0	196	138	0.704082	
197	1	197	139	0.705584	
198	1	198	140	0.707071	
199	0	199	140	0.703518	
200	1	200	141	0.705	

Figure 4.16. S’s sample size plot in TinkerPlots™ for Growing Sample Means: 0-1-1. The sampler, lower left, shows 0 with 33% and 1 with 67% chosen, and set up to add one value incrementally on to the sample; the sample, upper left, shows the value drawn, the sample size, the running total of 1s drawn so far, and the running mean so far; the sample size plot, right, shows the mean (x-axis) against the sample size (y-axis).

After S completed the simulation (Figure 4.16), they noted that the not only the center but the variability differed between the 0-1 and 0-1-1 sample size plots:

> Um, it didn't really zigzag any farther than, like, 0.7. [OK] Um, I think partially it was because of how many, um—we got so many blue blocks right in the very beginning, [mh] so beca—it would—it like—it's going to va—like, there's really—like for this one, for the one, um, before with only two blocks—um, the next one was orange, so, like, it would—it'd like—drastically changed, [Hm] while here, it's like, less of a sample size, like, once I did get a blue—once we did get a blue block, it didn't change quite as much. [OK] So there wasn't really as much as zigzagging, but it definitely pretty quickly it went, um, to the point s—like, around point sevenish. (2.903–918)

S started by noting that the range is restricted to greater than 0.7 and, without prompting, dug in to the individual outcomes to understand why this occurred. They recalled that in the 0-1 situation, the first block was blue, and the second block was orange (Figure 4.13), and so there was a large change. Here, in contrast, there were “so many blue blocks right in the very beginning”—the first 16 blocks were blue (Figure 4.16). Although S then says that “it’s, like, *less* of a sample size” (emphasis added) and that “once we did get a *blue* block, it didn’t change as much” (emphasis added), I interpret them to be commenting that in this case, by the time an *orange* block came along, because there was already a *larger* sample size at that point, the mean “didn’t change quite as much” due to the larger sample size.

This interpretation is consistent with their concluding comment on the 0-1-1 situation when asked what would happen if they repeated the simulation:

> Um, I think depend—if we—like a couple of orange blocks in the beginning, um, I'd say it'd be more zigzaggy most likely I would say, but there's also like h—for this one, it's like a higher percentage of blues compared to the oranges so it makes sense, how they're like—it's gonna be a little bit more towards like the blue side, so. (2.953–962)

At first, S noted that more orange blocks at the beginning would lead to more zigzagging, but then they noted that more orange blocks would be unlikely to happen because the population percentage of blues was higher. This meant that there would be less of a mixture of oranges and blues at the beginning, and thus less zigzagging in the 0-1-1 situation. Putting these pieces together adds up to a quite sophisticated chaining of the mechanism of sampling variability that links the population setup to the variability: There was a higher chance of drawing blue, therefore it was more likely to have fewer oranges early on in the sample. Therefore, there would be less drastic changes in the mean, because the few oranges will not have as large an effect on the mean, which is because there will be enough of a sample size including many blues already. This chaining throughout the mechanism links the population sampling process to swamping for the first time. Previously, S had described swamping but with no connection to sampling, simply as a property of means and proportions, and then separately commented on the long–run average, whereas here S has coordinated swamping with the stochastic process to reason about the situation to explain the differences in variability. S appeared to draw extensively on the representation of "zigzagging" in the sample size plot, as well as the ability to map movements in the mean directly to what each sample outcome was, since when a 0 was drawn, the mean moved left, and when a 1 was drawn, the mean moved right, and S could track the correspondence between each case and its influence on the mean. This may have also supported S's attention to streaks of 0s and 1s, which was crucial to the above chain of reasoning. Such streaks may be more salient also because S was physically clicking the RUN button in the TinkerPlots™ simulation (Figure 4.16) for every individual value and observing the changes happen over time. Even after this complex reasoning and simulation experience, however, S still did not clearly link the population mean to the mean of a large sample.

4.3.5 Growing Sample Means: Cat Factory 1

The Cat Factory activity stretched S even further from their comfort zone of coin–flip situations by allowing many more possible values: the integers between 6 and 32. Because S was encouraged to set the TinkerPlots™ sampler to what they thought was reasonable, S attended to properties of the population such as *mean*, *mode*, and *distribution*, as well as the fact that the *sampling* was random (Figure 4.17, left). Throughout this segment, S predicted and observed that the mean was not varying or zigzagging as much as in the dichotomous situations, thus the activity *Sample mean stays.* S also reasoned about causal relationships between the population mode and the sample mean (Figure 4.17, right).

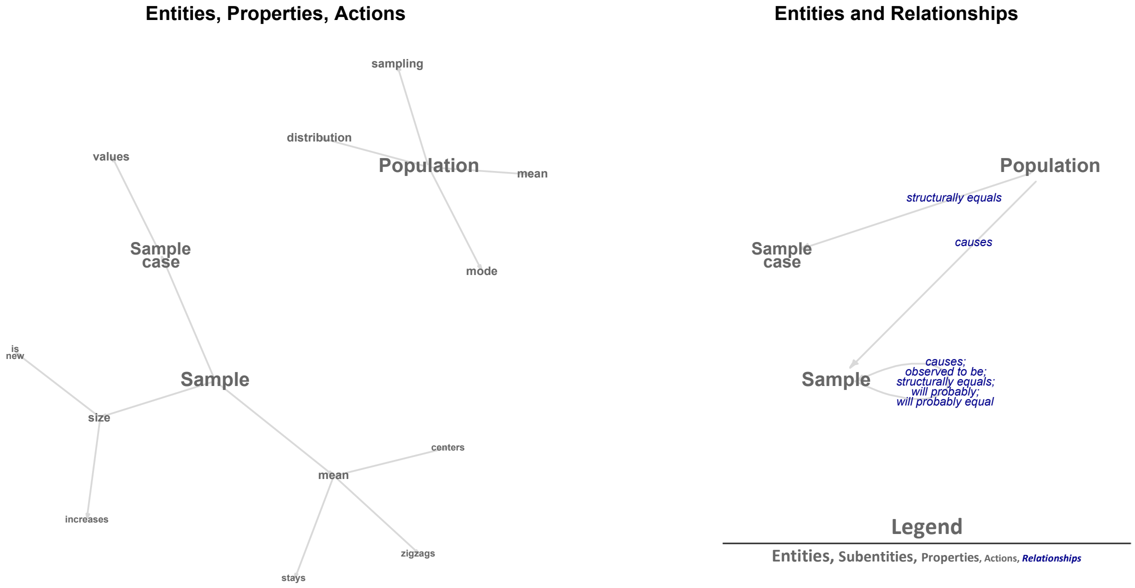


Figure 4.17. Summary of coded entities, properties and actions (left) and entities and relationships (right) for the Growing Sample Means: Cat Factory 1 activity for participant S.

S drew a bell-shaped curve with the mode near 20 inches for cat length production (Figure 4.18). Their reasoning about what would happen as the sample size grows seemed to rely largely on the mode:

> Um, I guess I think at the very end it'll probably be, like, around like the 20ish area, kind of like where it started, just because, like, it's more likely to have, like, a cat that's, like, the average—like, say, 20 or something like that, than it is to have, like, a cat that's like 32 inches or something. (2.1031–1039)

Although they mention the population mean, their reasoning depends on what values are most likely in the population. Values in the sample are most likely to be near 20, and so the mean is likely to be near 20 as well. As in the Post Office Simulation activity, this reasoning led to correct predictions in this case because the mode and the population average are the same, but S would soon experience a population where the modes and mean were not aligned in Growing Sample Means: Cat Factory 2.

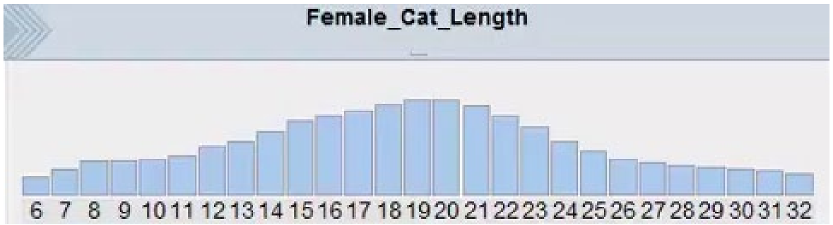


Figure 4.18. S's TinkerPlots™ sampler setup for Growing Sample Means: Cat Factory 1

S was convinced that there would be less zigzagging in this case than there would be in the dichotomous situations. The displayed TinkerPlots™ setup may have given some clue that this may be the case, because the sample size plot for the first five means was displayed for S and showed relatively little variation already. S reasoned that this was both because of the kind of population and the properties of the average:

> Well, in, like, say, like, the 1 to 2—or like 1 to 0, it can't really be anything in between, it can only be one thing or another. [Hm] While this one it can be, like, multiple different things, [mh] so it won't be zigzagging quite as much, like, all around, it'll be more of, like, um, kinda staying in the middle. And also it's taking the average—or it—like the—the me—it's, like the mean, so it's not going to be, like, varying too much, because it's taking that mean of everything, so. (2.1055–1066)

S appeared to be concluding from the fact that the population has more possible values that the mean will not zigzag as much. There may be some correct reasoning here. Although it is difficult to normatively compare the amount of zigzagging between this Cat Factory sampler and the previous dichotomous situations, the sample size plots so far have been set to the minimum and the maximum of the possible values. By definition, a dichotomous variable would therefore have all its probability mass at the extremes, maximizing the variability given the possible values, whereas cat lengths have values in between the extremes. Clearly not all this reasoning was in S's response: If the population values were heavily distributed towards the extremes, it would not be "staying in the middle" as S described. S also attributed the reduced zigzagging to the fact that "it's taking the mean of everything". It is not clear quite what S meant by this, but there may be some sense of swamping in their statement, i.e. that the influence of each case was moderated by all the other cases. However, S's statement was not unique to means, since the percentage blue was also the "percentage of everything", so it is not clear what distinction S was trying to make.

4.3.6 Growing Sample Means: Cat Factory 2

Much of S's reasoning in Cat Factory 1 led them to correct inferences because the mode and mean of a bell-shaped distribution are the same. However, crucial complexities and gaps in S's reasoning about the mechanism of sampling variability were revealed when

S was given an opportunity to manipulate the mechanism—to change the sampler in a way that would produce a “different” sample size plot.

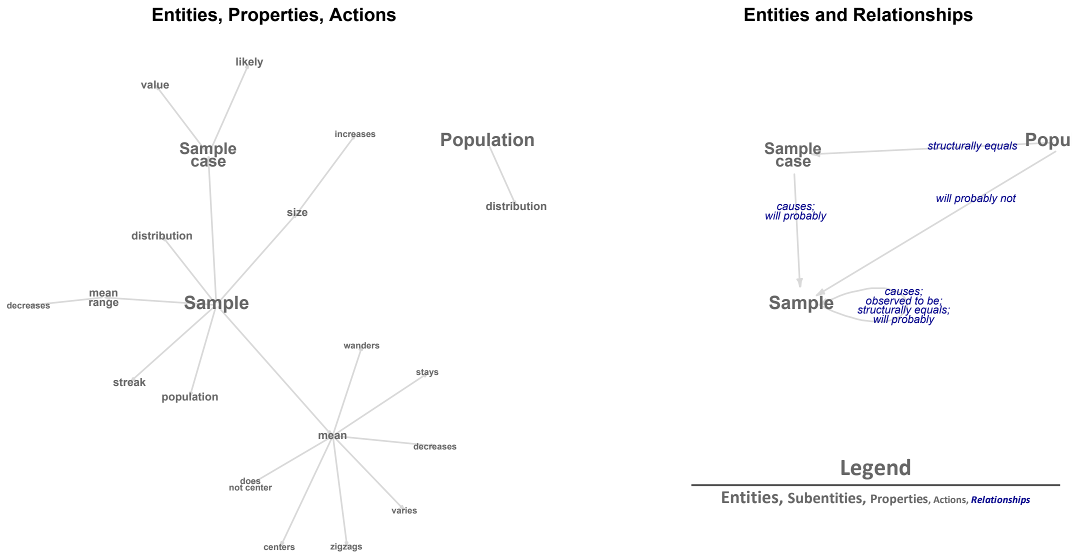


Figure 4.19. Summary of coded entities, properties and actions (left) and entities and relationships (right) for the Growing Sample Means: Cat Factory 2 activity for participant S.

A number of new codes appeared in this segment, perhaps because of this new opportunity to manipulate the mechanism (Figure 4.19, left). S chose a uniform distribution for the population, and focused on how the extreme values would be more structurally likely (Figure 4.19, right), and the presence of the extreme values of the sample would lead the mean to not be centered. This reasoning yielded the new code of *Sample case likelihood* here, with lots of focus on unusual activities of the *Sample mean* since S predicted that it could *wander*. S justified the choice of the uniform distribution as producing a different sample size plot, again with a focus on the mode of the population:

> I think it would zigzag more, um, just because there's like—a more likely chance that, like, each o—e—like, different lengths are gonna be drawn. [OK] Rather than, like, right now it's saying that, like, 20 is, like—gonna most likely happen, while if they were more even, like, it'll be more likely, like, say that, like, you could have the 30 inch cat [Hm]. Um, so. (2.1154–1163)

S appeared to be thinking that there will be more variability in the sample because now all the possible values are equally likely, and therefore there will be more zigzagging in the mean in the sample size plot. This is accurate; however, S may have been thinking that having all values equally likely is maximizing the variability of the population distribution, a misconception that delMas and Liu (2005) has found in other statistics students.

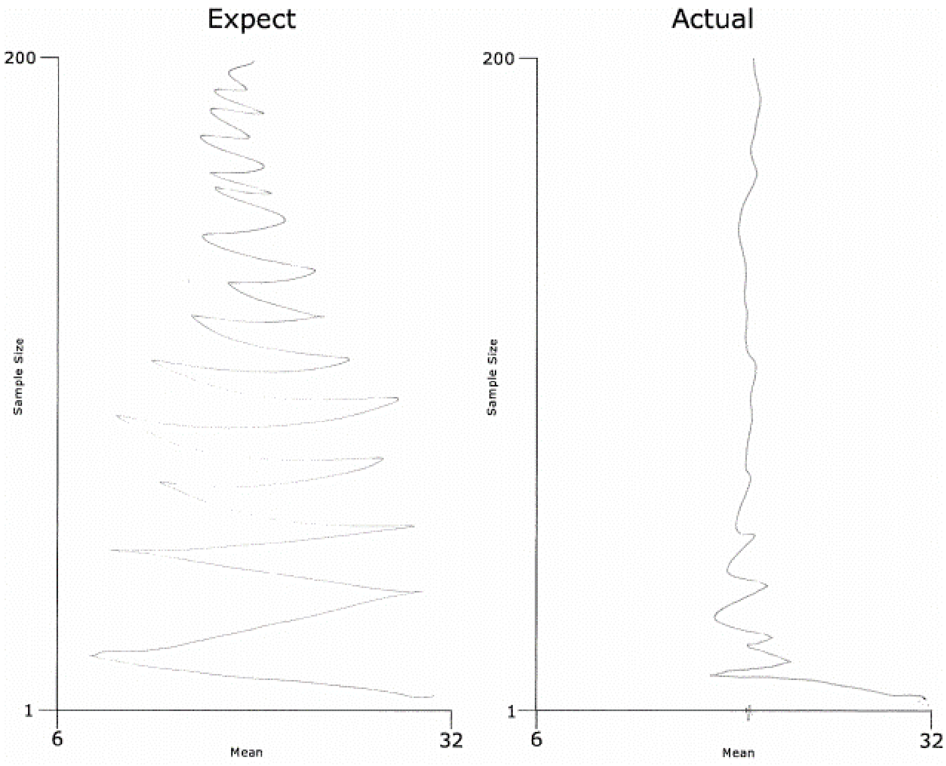


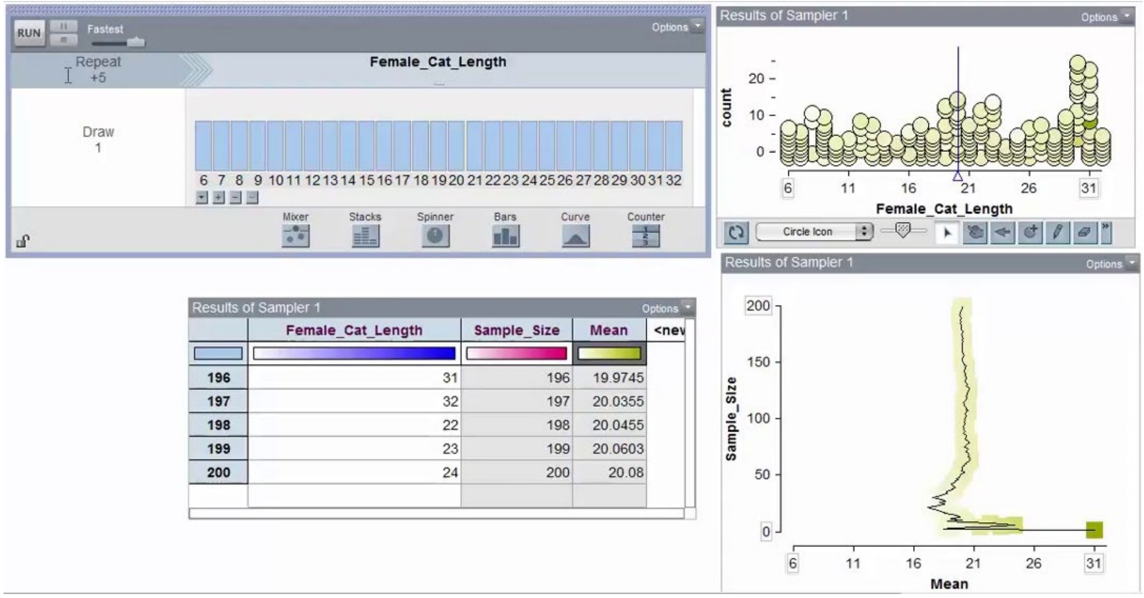
Figure 4.20. S's predicted (left) and observed (right) sample size plots for Growing Sample Means: Cat Factory 2.

S drew a relatively normative sample size plot (Figure 4.20). Although they overestimated the variability and the expected spikiness, they correctly drew the decreasing zigzags as the sample size increased and the eventual long-run center at the midrange, which is the population mean. However, when describing the graph, it became clear that S thought that the mean might not ever settle down:

> S: I think the mean can kind of vary cuz if every single chance is, like, likely, like, you could potentially have, like, a lot of runs that you get, like, a really large cat, compared [Hm] to like a really small, just for like the—trial. So I think the mean could definitely, like—it'll var—I think it'll—it'll vary, and it probably won't be centered in one area—area, maybe. [OK]

> I: So say that again. It'll be—it'll vary and it won't be centered in one area?
>
> S: Like—while, like, the other ones—they, like, were usually, like, centered around, like, say, like, say 20, or something, like, it won't really have, like, one spot that'll probably, like, kind of stay in potentially. Like, [OK] there could be like a quite a few rounds that where, like, the average is like more like on there like towards thir—32 but it could also like go back to a different spot. [OK] So. (2.1205–1230)

While S's conclusions are not normative, S displayed a lot of rich chaining about the mechanism of sampling variability. Again, S attends to the uniform population distribution of having the special property that "every single chance is likely". This unusual property created the possibility of having "a lot of runs". This vague reference to "runs" likely means streaks of large or small cats. S later extends their reasoning about runs after seeing the simulated sample size plot, and this attention to streaks of values is consistent with their reasoning on prior tasks, especially Growing Sample Means: 0-1-1. According to S, these large streaks of values then allow the mean to not converge to any particular value because there could be a streak of values that are all at 32. Therefore, the mean may stay there for a while, and when another streak occurs near some other value the mean then "could also go like back to a different spot". S was correctly coordinating many different parts of the sampling process, they just failed to recognize that streaks that could lead to a wandering mean are extremely unlikely. S still recognized that the "zigzag will get smaller" (2.1237), but decided that "it's not, like, afraid to move around" (2.1261). This clearly demonstrates the inadequacy of swamping alone as a mechanism for explaining the Empirical Law of Large Numbers: S recognized and fully incorporated the swamping in this situation without predicting the eventual convergence to the long-run mean.



Results of Sampler 1

	Female_Cat_Length	Sample_Size	Mean	<nev
196	31	196	19.9745	
197	32	197	20.0355	
198	22	198	20.0455	
199	23	199	20.0603	
200	24	200	20.08	

Figure 4.21. S’s sample size plot in TinkerPlots™ for Growing Sample Means: Cat Factory 2. The sampler, upper left, shows a uniform distribution, and is set up to add five values incrementally on to the sample; the sample, upper right and lower left, shows the value drawn, the sample size, the running total of 1s drawn so far, and the running mean so far; the sample size plot, lower right, shows the mean (x-axis) against the sample size (y-axis).

After seeing that there was, in fact, convergence to a single mean (Figure 4.21), S adapted their reasoning about streaks to explain this result:

> I: So what do you notice?
>
> S: It centered around, like, 21ish. [OK] Yeah. So it did center. But—and it didn't really deviate once, like, the—like it never really went towards like the six area. It kind of just like stayed higher. Like, it started out 31, and then it went down, and then it kind of just like kept on like staying like near the 21.
>
> I: Okay. And so why do you think that would be?
>
> S: Um, I guess—it looks like in the very beginning, there was probably more of the higher numbers for the mean, maybe. Um, and so then, it kind of, like, made it be more around that number. But it kind of—I don't think there's a lot of—like, ju—say like in this one,

> like, I don't think there was a lot of the lower numbers that, like, changed it quite as much in the beginning. [Hm] So by the time, like, say, like, there were lower numbers, then, like, the—I guess like the average between the two. So like if you have, like, in like the sample of, like, the five or whatever, it'll be, like, 31 and six, like, the middle of that is probably going to be more towards like the twenty area, say, than, like, centered around somewhere else, so. (2.1265–1300)

S observed that the sample did center, but also that it never went far below the midway point. In explaining this, S again provided a complex chain of reasoning involving streaks of extreme values. Noting that the mean stayed high, S concluded that there was a streak of higher numbers early on. Then, S noted that if there were already enough higher numbers, lower numbers would not actually drag the mean all the way to a lower extreme as S had previously predicted. Instead, with a combination of high and low numbers, "the middle of that" was going to be in the middle of the overall distribution. Since S already had a relatively well-developed chain of causation from the distribution all the way to the sample size plot, they could add the concept of *balancing* (Well et al., 1990) between the high and low extreme values which ultimately led to a mean in the middle. This combination of attention to the population distribution, swamping, and balancing formed a relatively coherent account of why the mean ended up where it did.

S's original reasoning about the population mode in combination with swamping worked in the bell-shaped situation, but broke down in the uniform situation. The addition of balancing now led to a coherent and correct explanation, but balancing (at least as expressed by S) is not sufficient as a mechanism for non-symmetric distributions. Given S's ability to adapt and extend their reasoning when confronted with a uniform distribution, it would be interesting to see how S would react to a non-symmetric non-dichotomous distribution, but Growing Certain did not provide any opportunities to explore how S would have reacted to this.

S reiterated that the zigzagging was larger on the bottom and tapered off at the top, and also noted that the uniform population distribution "centered a—much later" (2.1320). S recalled that in prior graphs "it was around like twenty or thirty it started kind of centering" (2.1328–1329), whereas with this one it was at 50 "where it started being more even and, like, less zigzaggy" (2.1332–1334). This is another point where S attendedto the point at which the mean starts to even out, which they first attended to in Growing Sample Means: 0-1 after seeing the sample size plot. The sample size plot appeared to give a useful visual for identifying that point of "evenness". However, the visual inference was apparently not particularly reliable: Relative to the population range, the sample size plot for Growing Sample Means: 0-1 (Figure 4.13) was clearly more zigzaggy at all sample sizes than was this sample size plot (Figure 4.21), despite S's memory of it.

Nevertheless, the sample size plot appeared to be a useful resource for S to both articulate in prediction, and explore in simulation, the dynamics of sampling variability. Although S's ability to apply the principles expressed in their responses in other situations in the later interviews proved to be limited, the depth of mechanistic reasoning shown in the responses above—coordinating sequences of numbers at different sample sizes, population shape, variability, and the convergence of the sample mean—was substantially more well-fleshed out here at the end of the second interview as compared to S's initial responses in Growing Sample Proportions: Physical Simulation, and there appeared to be a steady progression of building vocabulary and connection as the activities progressed. In particular, the ability to track the relationship between different outcomes at each sample size and the mean appeared to be particularly important for S. In both their reasoning above and in their concluding reasoning in Growing Sample Means: 0-1-1Growing Sample Means: 0-1, S attended to groups of values and how those affected the mean on the sample size plot, and chained backwards to the population process and forwards to the long-run impact on the mean.

4.4 The Mystery Mean & Growing More Means

This interview marked the transition from the sample size plot to the ESD, and therefore S reasoned frequently about the ESD towards the middle and end of this interview (Table 4.4). Revealingly, S was often reasoning about various elements of the mechanism from the perspective of an individual sample, and *Sample* codes were dominant throughout the interview. Also, while S did not reason much about the population mean during the Growing More Means activity, the Mystery Mean activity appeared to elicit more reasoning about the population mean.

Table 4.4
Presence of Entity, Property, and Action Codes that Occurred in More than One Section of Interview 3 for Participant S

Entity	Property	Action	Presence in Each Interview Section					
			MM1 One	MM1 Many	MM2 One	MM2 Many	GMM 0-1	GMM 0-1-1
ESD			_	■	_	■	■	■
ESD		varies	_	_	_	■	■	■
ESD	mode		_	_	_	_	■	■
ESD	modes		_	■	_	■	■	■
ESD	modes	combine	_	■	_	■	_	_
ESD	probability > 0.9		_	■	_	■	_	_
ESD	rank		_	_	_	_	■	■
ESD	shape		_	■	_	■	_	_
ESD column			_	_	_	_	■	■
ESD column	height		_	_	_	_	■	■
ESD column	number of permutations		_	_	_	_	■	■
ESD column	value		_	_	_	_	■	■
ESD sample			_	■	_	■	■	■
ESD sample	mean		_	■	_	■	_	_
ESD sample	size		_	■	_	■	■	■
ESD sample	size	increases	_	■	_	_	■	■
Population			■	■	■	■	■	■
Population	mean		■	■	■	■	_	_
Population case			_	■	_	■	_	■
Population case	value		_	■	_	■	_	■
Sample			■	■	■	■	■	■
Sample	mean		■	■	■	■	■	■
Sample	mean	changes	■	_	■	■	_	_
Sample	mean	jumps	_	_	■	■	_	_
Sample	mean	stays	■	_	■	_	_	_
Sample	mean	varies	■	_	_	■	_	_
Sample	probability		_	_	_	_	■	■
Sample	size		■	■	■	■	■	_
Sample	size	increases	■	■	■	■	_	_
Sample	values		■	_	_	_	■	_
Sample case			■	_	■	■	_	■
Sample case	value		_	_	■	■	_	■
Slot Machine			■	■	_	_	_	_

Note. A symbol of _ indicated that the code did not occur within that interview section, whereas ■ indicated that the interview section contained the code. MM = Mystery Machine, GMM = Growing More Means.

4.4.1 Mystery Machine #1: One Sample

The first mystery machine activity gave S an opportunity to show their reasoning about sample size in the situation of an unknown population whose true mean was quite far from the population mean. Many entities and relationships seen in the Growing Sample Proportions & Means session recurred in this section, including the zigzagging of the mean on the sample size plot at larger and smaller sample sizes, and attending to changes when a single case was added (Figure 4.22, left). Since S's task was to determine whether a given sample size was sufficient evidence to determine whether the machine's claim was accurate, new codes included evaluations of the population mean and whether there was *sufficient evidence* (Figure 4.22, right). S again referred to the number of games they imagined being played by a person in the casino as S attempted to bring in context (*Slot machine games per person*).

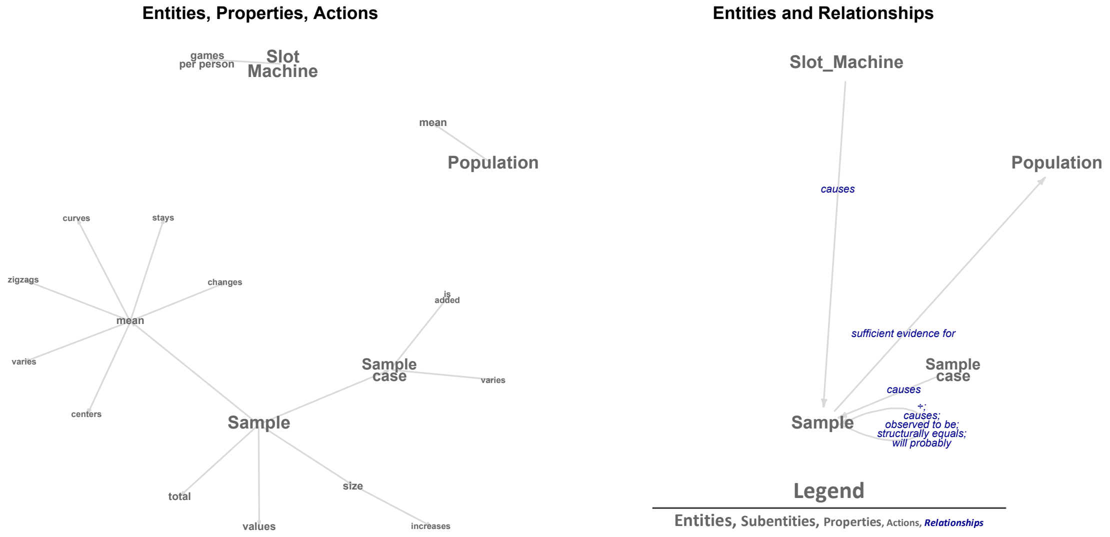


Figure 4.22. Summary of coded entities, properties and actions (left) and entities and relationships (right) for the Mystery Machine #1: One Sample activity for participant S.

S was initially asked what they thought the sample size plot would look like. They explicitly drew on their memories of the previous session, saying that they were "kind of thinking about last time" (3.73–74). They gave a specific example of how larger changes could happen in the bottom, building on their streak-based reasoning that they had developed in Growing Sample Means: 0-1-1Growing Sample Means: 0-1 and Growing Sample Means: Cat Factory 2:

> Like, if you say, like, win like $20 on your first time, like it's gonna be over here, but then if you win like zero time—or $0 the next time, like it's kind of gonna go zig—it's gonna zigzag more at, like, widely I guess. Like a bigger range, like two points, um, in the very be—like beginning, but then as you do more of the, um, slot machine then it should kind of at least the mean might go somewhere more towards, like, I guess 0, 1 or 20, not really sure which one but. /c/ [OK] Yeah. (3.74–3.86)

S attended to how widely the mean could vary at a very small sample size, giving the example of how a 20 and a 0 would cause a zigzag, and then commenting that the mean would eventually find a center. Note that at the beginning of Growing Sample Means: Cat Factory 2, S had been uncertain if the mean would ever center, but without knowing the population, S did expect it to center.

S then drew the predicted sample size plot, assuming the machine's claim was accurate (Figure 4.23, left). S expressed swamping a bit more clearly at this point, giving the 20 and 0 example again and noting that "if you've done it fifty times […] it's not—just not gonna affect the overall, like, result as much" (3.117–3.119). S was somewhat surprised, however, by the eventual shape of the distribution (Figure 4.24):

> Um, it was more of like—it's more a, like, curve than like zigzagging all the time. [OK] Um, I think it's probably because it's taking like the mean each time, so like it's not gonna vary quite as much, say, than, like, each individual like time you go. [OK] So. (3.202–208)

S here attended to how the change in the distribution was smoother than prior examples, especially compared to the sample size plots of dichotomous variables in the previous session. S attributed this to the action of taking the mean and appeared to be saying that the mean was not as variable as each individual case. This was an echo of their reasoning in the second interview about how the mean doesn't move as much because "it's taking that mean of everything" (2.1066), implying that the mean includes more and therefore moves around less than individual cases.

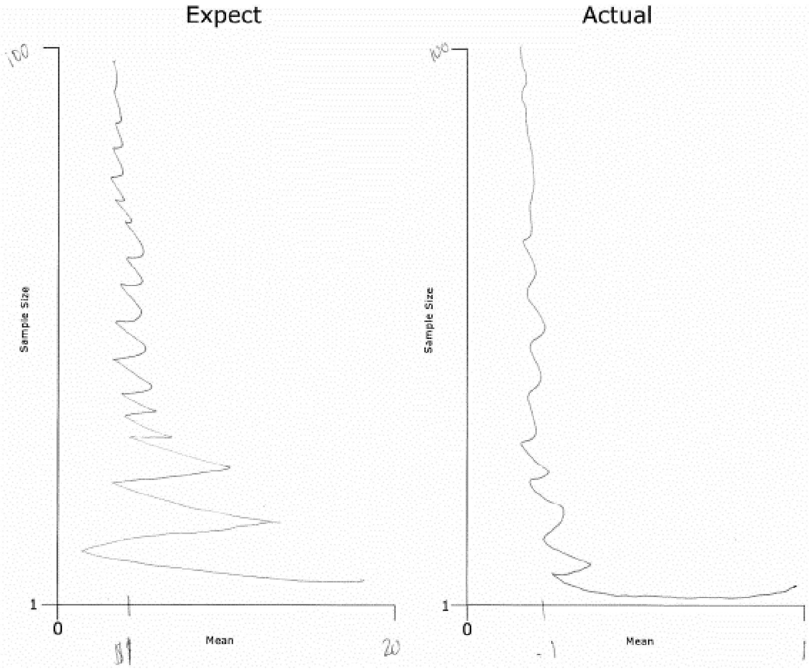
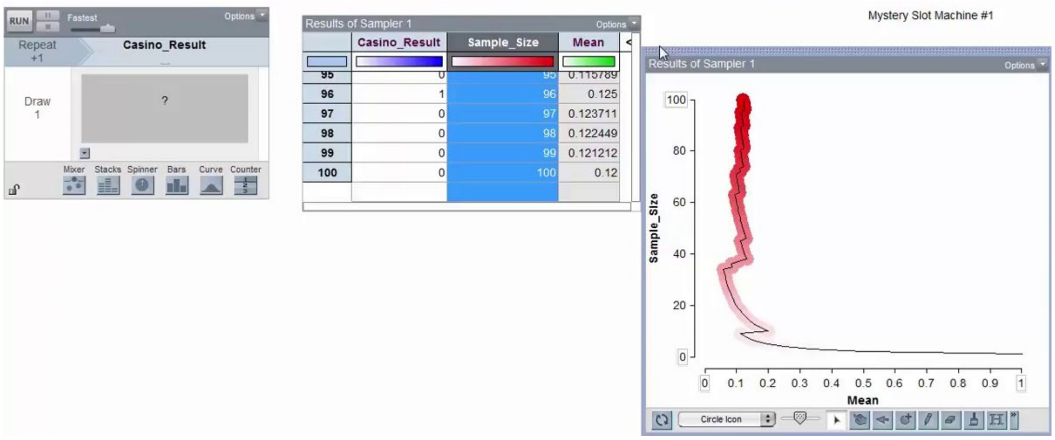


Figure 4.23. S’s predicted (left) and observed (right) sample size plot for Mystery Machine #1: One Sample.



Results of Sampler 1

	Casino_Result	Sample_Size	Mean
95	0	95	0.115789
96	1	96	0.125
97	0	97	0.123711
98	0	98	0.122449
99	0	99	0.121212
100	0	100	0.12

Figure 4.24. TinkerPlots™ setup for Mystery Machine #1: One Sample for participant S. Left, sampler with hidden contents; middle, case table calculating cumulative mean and sample size; right, sample size plot.

S expressed a high level of confidence that the machine's claim was inaccurate, because the mean went down so quickly to the 10–20 range and stayed there. When answering a probe about whether they thought they had enough evidence, S unexpectedly brought in their reasoning about the context of the slot machine:

> I: Um, okay. So based off of this you're thinking that it's inaccurate. [mh] Um, how comfortable do you feel with that? Do you feel pretty confident? Do you feel like you have good evidence that it's inaccurate? Or do you feel like you need more evidence?
>
> S: I'd say just doing, like, a hundred trials, I'd say I'd be, like, comfortable saying that it's probably not accurate. Cuz if most—most people if they play they're only gonna pay like—play like three times, or something like that. Like, they're not gonna be playing probably like a hundred times in a row. But in general I think that it's probably inaccurate. /c/ (3.243–252)

Note that although the probe only mentions "evidence", S immediately discusses sample size. However, the reasoning for choosing a large sample size was not because of the decrease in variability that S has noted many previous times. Instead, S compared their sample size of 100 with what they regarded as typical for the number of games an individual person played at the casino. Since the sample size of 100 was much larger than the two or three games they expected a person to play, they concluded that 100 should be enough to tell if the machine's claim was accurate.

4.4.2 Mystery Machine #1: Many Samples

After exploring the growth of a single sample, the activity then bridged to examining the ESD of many samples. There was a high density of both entities and properties in this segment (Figure 4.25, left). Some of this profusion was due to the complexity of reasoning about samples, populations, context, and evidence simultaneously. Because of the complex shape of the ESD, S reasoned about modes, and, at the interviewer's prompting, about high values particularly those that were above the hypothesized mean of 0.9. Another contributing factor was the fact that S's ambiguous use of language necessitated generating new codes. For instance, S often used the word "trials" in an ambiguous way that seemed to both describe *ESD Sample size* and *ESD size*. In the CATALST curriculum, "trial" refers to the number of simulated samples (Zieffler & Catalysts for Change, 2017), or *ESD size*, but S's usage was ambigously broader, hence the new property *ESD number of trials*. S reasoned about causal relationships between the population and the sample, and in turn whether there was sufficient evidence to make a decision about the population, as well as commenting on the percieved structural relationships between cases in the ESD, the sample mean, and how the sample size aligned with the number of people they would expect to play the slot machine (Figure 4.25, right).

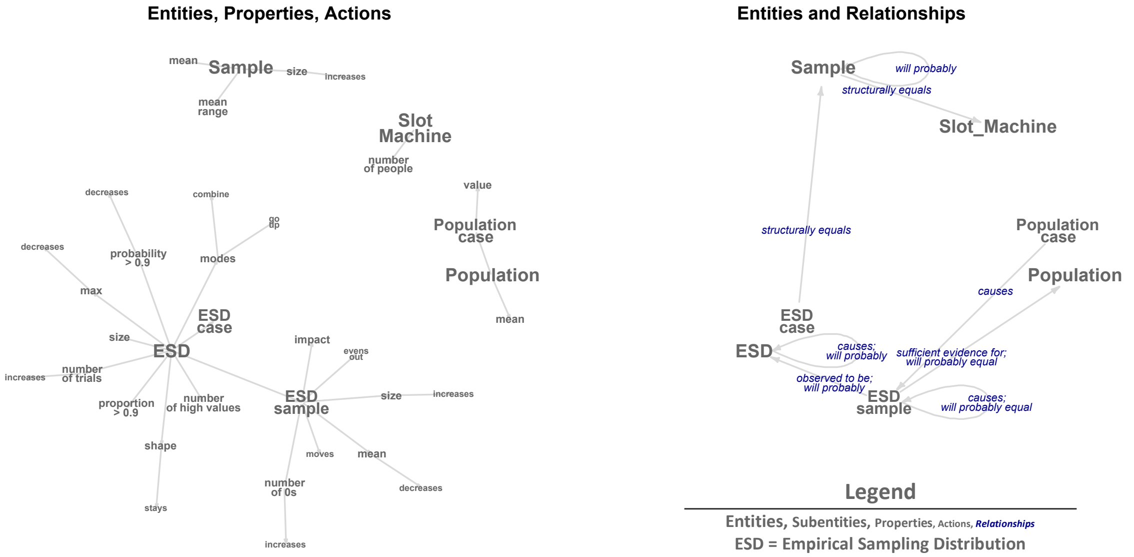


Figure 4.25. Summary of coded entities, properties and actions (left) and entities and relationships (right) for the Mystery Machine #1: Many Samples activity for participant S.

Once S understood that they were being asked about what they would expect for an ESD of 100 trials, they readily drew a plausible unimodal ESD (Figure 4.26, left), perhaps informed by their exposure to ESDs in the CATALST curriculum. The actual, bimodal result (faithfully reproduced by S in Figure 4.26, right) surprised them, however, since CATALST generally did not expose students to population distributions as ill-behaved as these Mystery Machines. When S was asked if they would be comfortable concluding that the machine's claim of 90 cents was inaccurate, their reasoning showed mixed evidence of understanding the different elements of the mechanism:

> S: Um, I would think so. I mean, each one of those dots represents the mean of 100 trials [mh], so there's a lot of trials on here and the fact that none of them—there was only one above, like, point, like, say six five—it's really like probably—even been playing it 20 times, like, you probably wouldn't even maybe get to, like, point 9 [OK], so. [OK] Yeah.

> I: Um, so how could you be even more sure that the, um—even more kind of confident that you have the right—that the machine is wrong, to gather even more evidence.
>
> S: Probably just running more trials, I would say. [OK] There's not—it never hurts to have more trials rather than less, so.
>
> I: Okay. And so, uh, how many—um, how many more, or...
>
> S: Um, I'd probably keep it at like 500 for like collecting the statistic, but maybe for like the draw on the original sampler, uh, maybe increase that to, like, 200 or 300 or something like that. [OK] Um, so. (3.433–463)

S here revealed their fluid use of the term "trials". S started out noting that each dot "represents the mean of 100 trials", and then said that "there's a lot of trials on here", apparently now referring to the means in the ESD since "there was only one" above 0.65. When asked about how to gather more evidence, S suggested "running more trials". But when asked about how many "trials", S felt the need to distinguish the two kinds of "trials" by referring to their locations within TinkerPlots™—*ESD size*, the "500 for like collecting the statistic" should stay the same, while *ESD Sample size*, the "draw on the original sampler" should increase. Although S's essential reasoning was entirely correct in this segment, S frequently had confusion about the distinction between *ESD size* and *ESD Sample size*, and it may be that S's undifferentiated vocabulary could be either reflecting or facilitating their conceptual confusion.

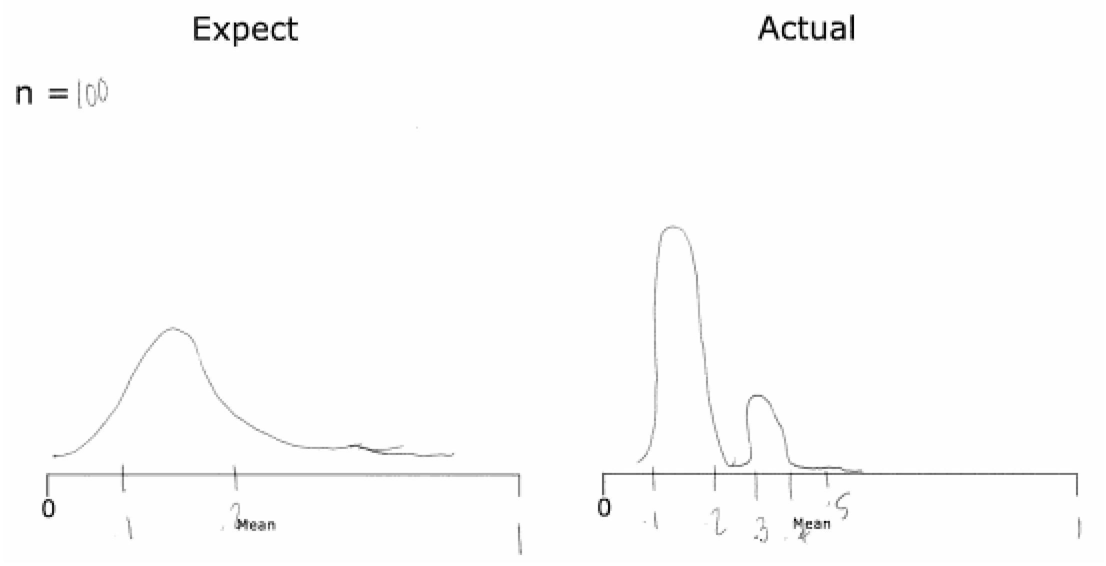


Figure 4.26. S’s predicted (left) and observed (right) empirical sampling distributions for S’s chosen sample size of 100.

This conceptual confusion was displayed clearly when S drew what they expected to see for the ESD of Mystery Machine #1 at sample size 200 (Figure 4.27a). Despite the fact that they had chosen this sample size to gather more evidence about whether the machine’s claim was accurate, they predicted that the sampling distribution would be the same and they drew a very similar predicted graph to the actual graph for sample size 100 (Figure 4.26):

> S: Um, I feel like it'll probably stay pretty close to what we had before, um, so we'll say, like, this is point five, and it'll kind of go up and then up again and then probably go off one more time, so this is like point two, um, point three, and point four.
>
> I: Okay. And why would you expect that to happen?
>
> S: Um, I—I don't really expect that much of a difference. Just because, um, we had so many trials in the previous one as well. [mh] So I feel like if anything, um, it'll just have—be, like, larger like the—each point will be larger, probably. [OK]

> I: And so could you describe what you mean by each point being larger?
>
> S: Like each, um, like, bump in the graph or like when it goes up it'll just kind of go up even more because there's more like on—like there's more trials, so. [OK] (3.512–537)

S almost appeared to be speaking about totally separate different objects when they spoke about the means in the ESD as opposed to the means in an individual sample growing via the sample size plot. S may have been thinking that after 100, the range of the ESD will not change, because at 100 there was "so many trials in the previous one". However, they expected that the modes will "go up even more because there's more […] trials". This unclear statement was hard to map to any part of the mechanism. It is possible that S was positing that the means will cluster more in an ESD with a larger sample size *in a bimodal way*—the means would concentrate more around each mode—but did not draw the further conclusion that this "bimodal heaping" would still imply that the ESD for sample size 200 would be expected to have fewer sample means in an extreme area of the ESD. Another interpretation is that S thought that somehow there will also be *more means* on the plot, raising the whole plot overall, which is consistent with the confusion between *ESD size* and *ESD Sample size* that S had displayed in the first interview.

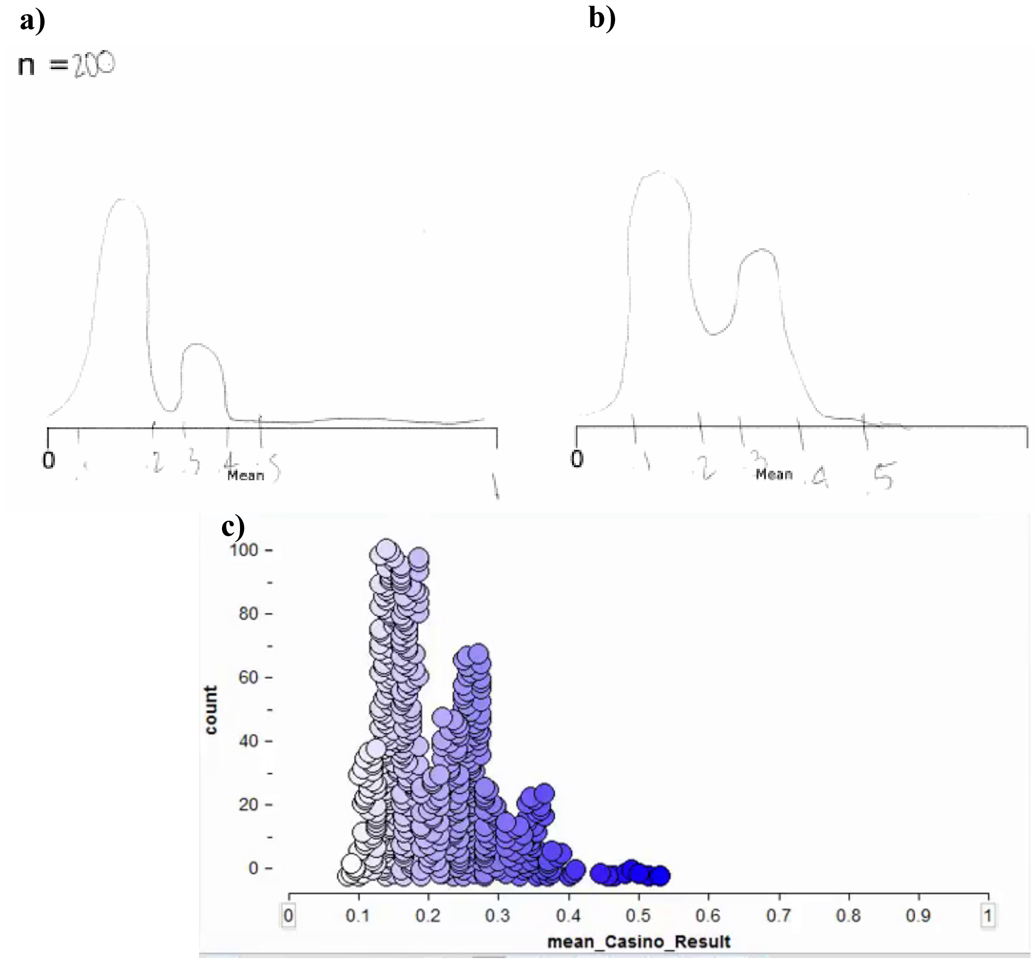


Figure 4.27. Predicted (a), sketch of observed (b), and TinkerPlots™ (c) empirical sampling distribution for S's second chosen sample size of 200 for Mystery Machine #1.

S noted that the maximum value decreased and thus that getting a mean of 90 cents would be even more unlikely when playing 200 games (Figure 4.27c). They provided a relatively sophisticated explanation about why a single sample would be less likely to be that high, because if $1 or $20 were unlikely, then when a hundred more are added, "that's more zeroes just adding into that, like into the mean, um, and so it's definitely gonna decrease” the mean for each trial (3.575–577). For the first time, S was able here to successfully reason about the mechanisms of what happens in a single sample to explain a result they saw in the sampling distribution, without the support of the sample size plot. However, when asked why the shape changed and the modes “combined” at sample size 200, S had more trouble coordinating the different pieces of the mechanism:

> S: Um, I'd say just having more trials and, like, it's—I'd say just like having more trials, like, it's kind of like just evening out I guess a little bit more. Um, [OK] and probably getting more centered—or like it's just kind of—yeah it's just evening out. /c/
>
> I: Okay. Can you tell me more about what you mean by “evening out”?
>
> S: Like it's gonna—like each dot isn't gonna be moving quite as much as like if you only have like say 20 trials. Like it doesn't make as much of an impact. [OK] So it's gonna be more likely to be like around what the average amount of pay—payout is like in general. [OK] So. (3.603–621)

The visual metaphor of “evening out” could make sense in the sample size plot—the line straightens and becomes more even as the sample size increases—but it is unclear how this applies to an ESD. It is even more unclear what could be meant by “each dot isn’t gonna be moving quite as much”, since this representation does not include dots moving at all, nor is it clear what entity would make an “impact” here. In the sample size plot, the mean dot “moved” as the sample size grew, and each case had less of an impact on the mean. S again may be reasoning based on the sample size plot, and not knowing how to map the entities and activities foregrounded in that representation to the ESD representation. Alternatively, it is possible that S was imagining a dynamic representation where all the sample means are growing simultaneously, which they would not experience until the Growing Many Means activity, but it is still unclear what the “impact” is. Intriguingly, S thought that the means will be near “the average amount out pay—payout […] in general”, a phrase that was the closest S had ever gotten to expressing a clear concept of *Population mean* in any context besides for a 0-1 variable. In S’s original engagement with the casino problem, they had difficulty conceptualizing what such an “average payout” would even mean, but here they seemed to think that there is some sensible concept of an average payout “in general”.

4.4.3 Mystery Machine #2: One Sample

Mystery Machine 2 was more likely to have 20s. This meant that the sample size plot was likely to have more visible leaps in the mean. In discussing the sample size plot for Mystery Machine #2, S attended to these *Sample mean jumps* (Figure 4.28, left), as well as noting the impact of a value of 20 on the changes in the mean as the sample size increased (Figure 4.28, right).

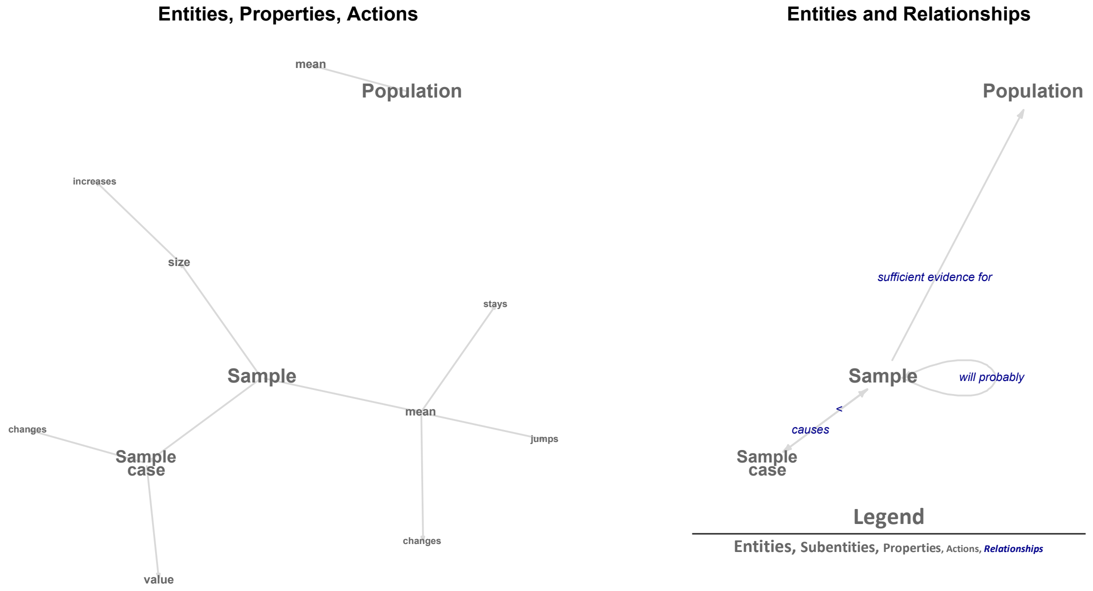


Figure 4.28. Summary of coded entities, properties and actions (left) and entities and relationships (right) for the Mystery Machine #2: One Sample activity for participant S.

S's attention was on the changes in the mean throughout this segment. Ironically, they drew even less variability for the mean for their updated prediction (Figure 4.29a), drawing on their prior reasoning that because they were taking the mean "it's not gonna be, like, super drastic of a difference, like, each time" (3.673–675), because the mean isn't going to change as much "each time you have, like, your own trial" (3.685–686), by which S probably meant that a case in the sample would move more than the mean that includes that case.

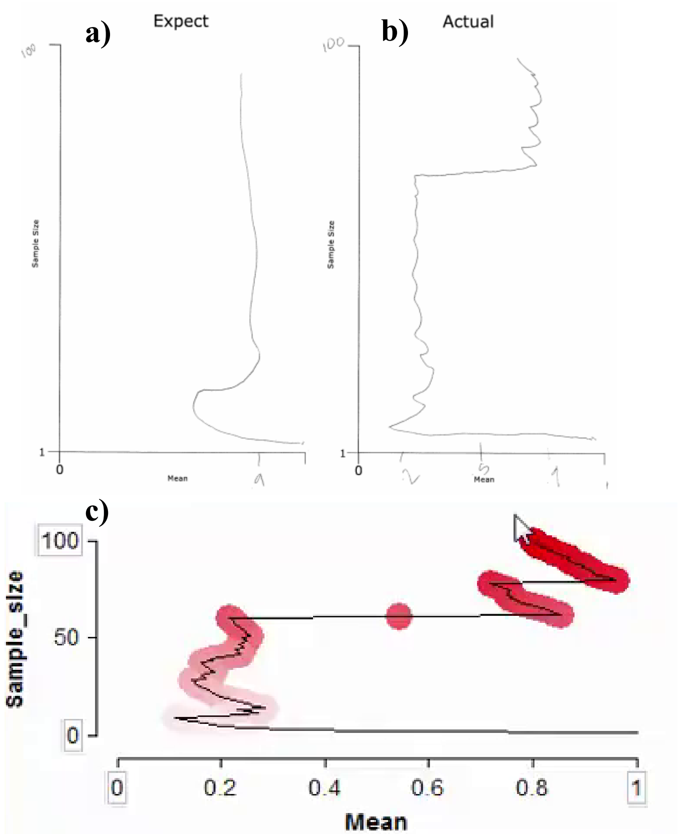


Figure 4.29. S's predicted (a), sketch of observed (b), and TinkerPlots™ sample size plot for Mystery Machine #2.

After creating the sample size plot in TinkerPlots™ (Figure 4.29c), S immediately commented "so that one definitely had a jump" (3.700), noting that the value of 20 must be responsible for the jump. When asked about whether they could tell whether the machine's claim is accurate, S responded *only* in terms of the variability of the mean, and not in terms of the fact that the actual value was near 0.9—perhaps because it was an obvious point, but also perhaps because the changes in the mean were so salient:

> I'd say no. You definitely would have to take a lot more samples. Because if one mean can change the plot of the graph, um, so drastically, I would say definitely like you would need many more samples /c/ or, um, trials at least to give like an accurate, um, claim on what it is, so. (3.735–741)

S used here the same word "drastic" that they used earlier during Growing Sample Means in the second interview. S was commenting that if one "mean" can change the plot so strongly, then they need more evidence in order to evaluate the machine's claim. S had noticed in Growing Sample Means the points at which the sample size plot stopped zigzagging and tapered into a straighter line; now, S appeared to be using the presence of tapering as a criterion to evaluate whether the mean is reliable enough in order to evaluate the machine's claim.

Again, however, S's use of terminology was confusing. S started by using "samples" to refer to additional cases to add on to the sample (e.g. *Sample case*), but then corrected their term to "trials". S also noted that one "mean" can change the plot quite a bit, which may be a sign that S was confused about what the plot represented. The mean did not cause the change—the new value caused the change. S also noted that when a 20 was drawn, "it, like, changed the mean for the later ones, too" (3.712–713); although it was true that the presence of a 20 would affect later values as well, the fact that S viewed that dependence as noteworthy may be a sign of confusion about how the individual values affect the mean. It may be that S simply misspoke, or it may be that they are not completely tracking that the sample size plot is showing the mean at each sample size. Therefore, a very extreme value by definition not only affected the plot at that one time but influenced the entire plot above (though less as sample size increases). The dependency of the upper part of the sample size plot on the lower parts of the sample size plot may be unclear to them and may be a drawback of this representation.

4.4.4 Mystery Machine #2: Many Samples

Mystery Machine #2 brought noticeable changes to the shape, location, and variability of the ESD as compared to Mystery Machine #1. Therefore, many codes tagged these changes in the properties of the ESD, especially the increased number of modes which S also referred to as "variability" (Figure 4.30, left). S traced these changes to the increasing likelihood of twenties. This section also challenged S to compare the two machines and determine a final sample size for the unknown machine, which led to varied reasoning about how the location and variability implied by the probability of $0, $1, and $20 payouts, and the resulting sample size needed to tell whether the machine's claim was accurate (Figure 4.30, right).

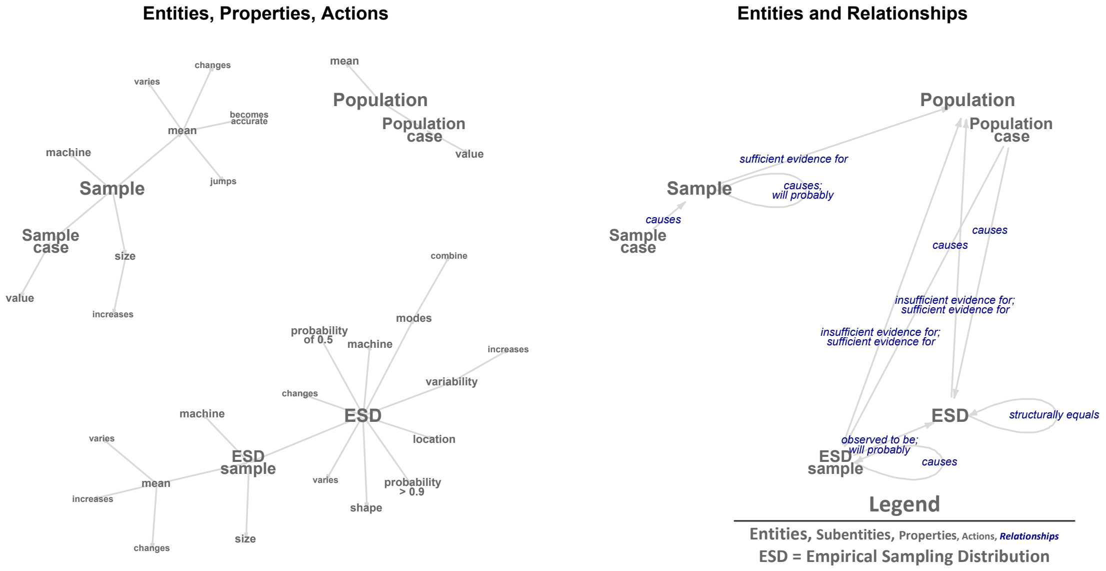


Figure 4.30. Summary of coded entities, properties and actions (left) and entities and relationships (right) for the Mystery Machine #2: Many Samples activity for participant S.

S did predict that the location of the ESD would shift to the right because of the higher probability of $20s, and that ESD would still have two modes (Figure 4.31a), but ultimately the distribution had more like four or five modes (Figure 4.31b–c). When asked why the ESD shape was different than Mystery Machine #1, S now said the higher percentage of $20s would make the mean "a lot more likely to vary" (3.896–897), again because the $20s will "affect the—uh, the mean, um, quite a bit" (3.904–905)—linking the observed variability to the potential for variability enabled by the mechanism.

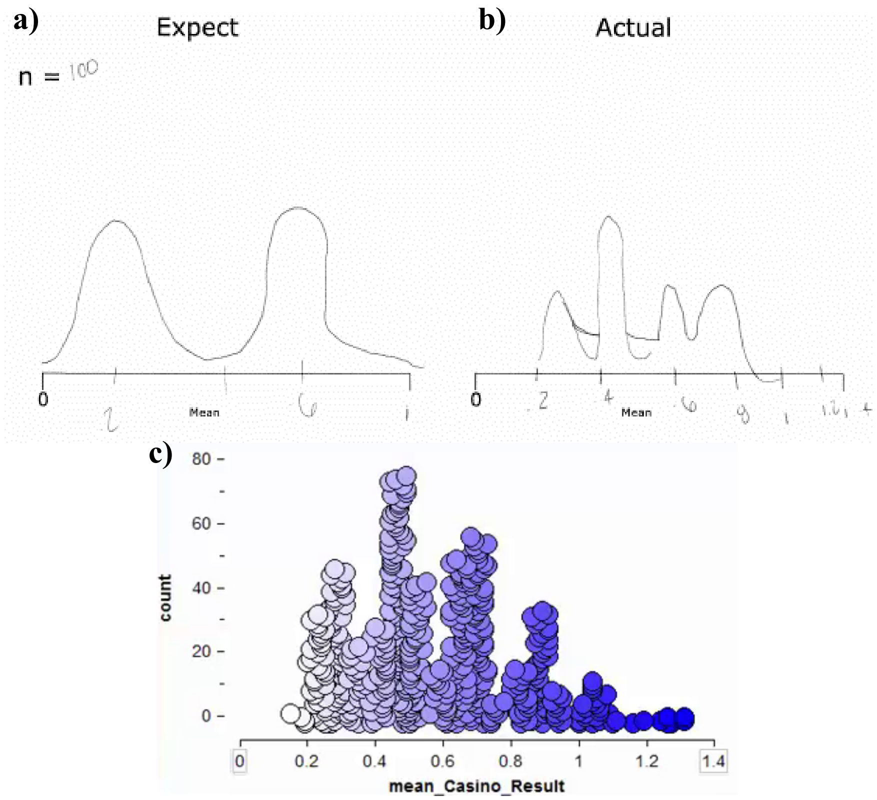


Figure 4.31. S’s predicted (a), sketch of observed (b), and TinkerPlots™ (c) empirical sampling distribution for Mystery Machine #2 with sample size 100.

As they did in Mystery Machine #1, S thought that they wouldn’t be able to determine whether the claim of 90 cents was accurate, because there was still too much movement in the mean. Even though S was explicitly asked what the probability was of ninety cents or higher at this sample size, and responded that “it’s, like, plausible” (3.922), their reasoning that the sample size was insufficient appeared to be entirely based on the sample size plot and not on the ESD:

> S: Um, I don't think so. I'm just thinking about like the previous graph, um, with like 100. Like you know it kind of goes back and forth quite a bit like even with like only a hundred

> trials. Um, so if like you s—get like, say if you do it like twice, like if you get a zero at one time and twenty dollars the next time your average payout would be ten dollars [mh] and like that's not really like accurate so you're definitely like I'd say you need more than like a hundred trials to be able to like say that the average would be ninety, but. (3.944–958)

S here referred to the "previous graph", presumably the sample size plot, and discussed how the mean "goes back and forth quite a bit". They again gave the example of drawing just a zero and twenty for the first two trials, and how far off the true mean that would be. Again, S seems to want the mean to not be moving as much in order to be able to make claims about the population average. The sample size plot appears to have made variability, and perhaps uncertainty, salient enough that S attended to that uncertainty even more than the location and the fact that the ESD has a fair number of values over the hypothesized mean of 90 cents; in contrast, prior research has generally found that location is generally more salient than variability to participants (Griffin & Tversky, 1992; Obrecht et al., 2007).

S predicted that the modes would combine more when increasing the sample size to 200 (Figure 4.32a), which in fact occurred (Figure 4.32b–c), coming together into "a volcano" (3.1004). S's reasoning about the shape revealed their non-standard usage of the term "variability":

> S: Um, it's definitely less variable, I would say. [OK] Or, it just like varies less, like there's less of it being like shooting up to like a certain mean or something, and it's just a lot more of like an average, um. I would say like it's—it kind of just like is a regular about curve, almost.
>
> I: Okay, okay.
>
> S: So. Instead of having like multiple ones.
>
> I: Okay. Um, and that's what you mean by varying? It doesn't have as many kind of—
>
> S: Yeah.

I: —separate peaks? OK.

S: Mm-hmm. (3.1024–1040)

S used the term “variability” to mean the tendency towards a multimodal distribution, “shooting up to like a certain mean or something”. This interpretation also provides more context to their description of the multimodal shape of Mystery Machine #2 at sample size 100, above—they attributed the shape to the percentage of 20s making it “a lot more likely to vary” (3.896–897). Some of this confusion about variability may be difficulty in translating different representations. The sample size plot shows an evolving process, and the “zigzags” of the mean do represent variability; in contrast, the ESD is the result of a repeated process (also interpreted as a probability of certain means), and does not directly depict the process of the growing sample . Many bumps are kind of like “zigzags” but instead of representing the movements of a single mean as sample size grows, the bumps represent groups of sample means at a given sample size. S’s language even suggests a kind of process unfolding as one moves from left to right across the graph, with the vivid language of “shooting up”; although this may simply be a metaphor to describe the shape, it may also represent subtle conceptual confusions about what the ESD represents.

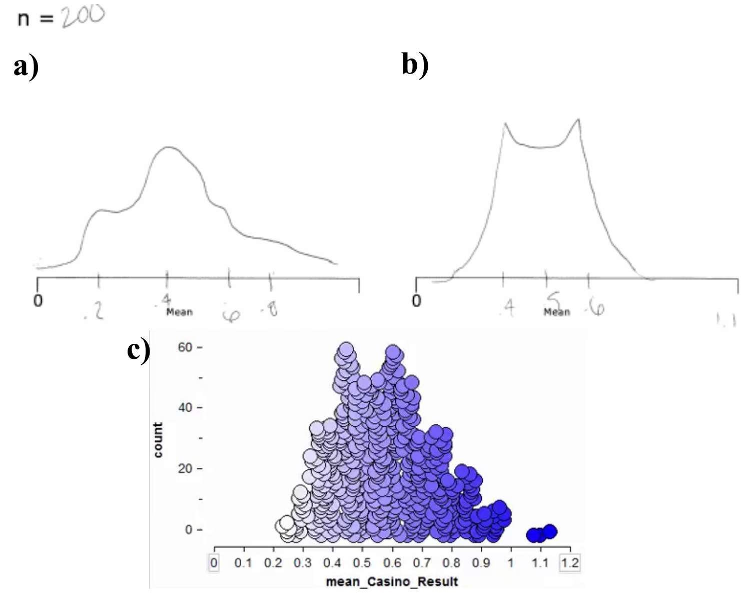


Figure 4.32. S's predicted (a), sketch of observed (b), and TinkerPlots™ (c) empirical sampling distribution for Mystery Machine #2 with sample size 200.

When asked whether they had enough information to make a claim, however, S immediately switched to using "vary" in an apparently normative way, expressing skepticism about the claims they can make and noting that "it still does vary quite a bit, um—there's—I mean it goes from around, like point two to point nine-ish" (3.1062–1065). Here the mean "varies" in the sense of the range of the ESD, which was both accurate and relevant here. S here reasoned for the first time about the variability of the mean as expressed by the ESD and not by the movement of the mean in the sample size plot. Additionally, S displayed some awareness that the location mattered, noting that Mystery Machine #2 was harder to tell whether it was accurate than Mystery Machine #1, because Mystery Machine #1 was "drastically lower than, um, what this one was" (3.1104–1105).

However, this appeared to be the extent of S's reasoning about the location of the ESD; S did not investigate the mean or median of the ESD either visually or by using the tools provided by TinkerPlots™. The only central tendency that S attended to was the modes, similar to how they generally only attended to population modes in the Cat Factory activities during the second interview. Although S could have paid attention to the mean of the ESD—the mean of the many sample means—in order to draw a conclusion about the population mean, the interview prompts discouraged this by noting that this mystery mean, and therefore this ESD, was only a hypothetical situation of what *could* happen and that the purpose was to determine how large a sample they would need to draw if they were to play the actual slot machine.

When asked how the casino owners could make it more difficult to tell whether the machine's claim was accurate or not, S fluidly mixed the two forms of "variability":

> S: I would probably say like making it more likely to get twenty dollars, because then it'll have a lot more like bumps and everything like that, so you won't really know like how likely it is, like—yeah, you could have an average payout of like say 20 cents, but you could also have an average payout of like $2, and like the—in between that like you're closer to like—say, like a dollar or something like that. (3.1120–1131)

Again S focused on the number of $20s by noting that the primary way to make the task harder would be to make the $20 outcome more likely. This would make add more multimodality, or "bumps", which would also contribute to uncertainty: "you won't really know how likely it is". They then gave an example of how they could get a range of different possible averages, which more hints at the range of the ESD, but both this range and the "bumps" do not appear to be differentiated concepts.

However, it is noteworthy that S’s reasoning was now more solidly based in the ESD rather than only the movement of the mean. S provided a relatively normative overall answer of the sample size they would need to have with an unknown population:

> I'd probably say at least having like—we'll say like five hundred trials. Because if you don't really know the—what like the probabilities are, like, it may work—like, only doing like a hundred may work for like one, but if it's like this one like—it—just doing 200 like it's a very very different graph, so I would say probably like doing like four hundred, five hundred, something like that, I feel like would give a much better result. (1148–1159)

They recognized and reasoned about the ESD effectively here, but even here, and certainly in prior quotes, they did not display much evidence of understanding what produced the ESD mechanistically. There was not a clear sense of how and why the ESDs were different, whereas S was able to articulate this relatively powerfully in terms of the sample size plot. However, the problem required reasoning about many possible means, and the ESD was the best tool available. The next activity explored S’s reasoning about ESDs and mechanism more deeply.

4.4.5 Growing More Means: 0-1

S now returned to the context of a 0-1 variable, with a focus now on sampling distributions and the likelihood of each possible mean. S immediately displayed heaping-like reasoning, but a lot of this reasoning was at first somewhat vague and about the possibilities of different types of samples, hence S spent a lot of time reasoning about *Sample values* (Figure 4.33, left). With the support of color-coding the sampling distribution by the permutation of values that created it, S was able to reify permuatations as its own subentity (*ESD permutations*), which they called “combinations”. S focused on several other subentities of ESDs as well, including the columns, *ESD column*, and the central region, *ESD central region*. S spent quite a bit of cognitive energy throughout thinking about what the possible values of the ESD would be, here coded as *ESD possible slots*. S fluidly reasoned back and forth about the likelihood of individual samples, and the heights of stacks on the ESD, leading to a quite complex web of relationships between entities (Figure 4.33, right) including causal relationships between the population, samples, ESD columns, ESD central regions, and the ESD permutations.

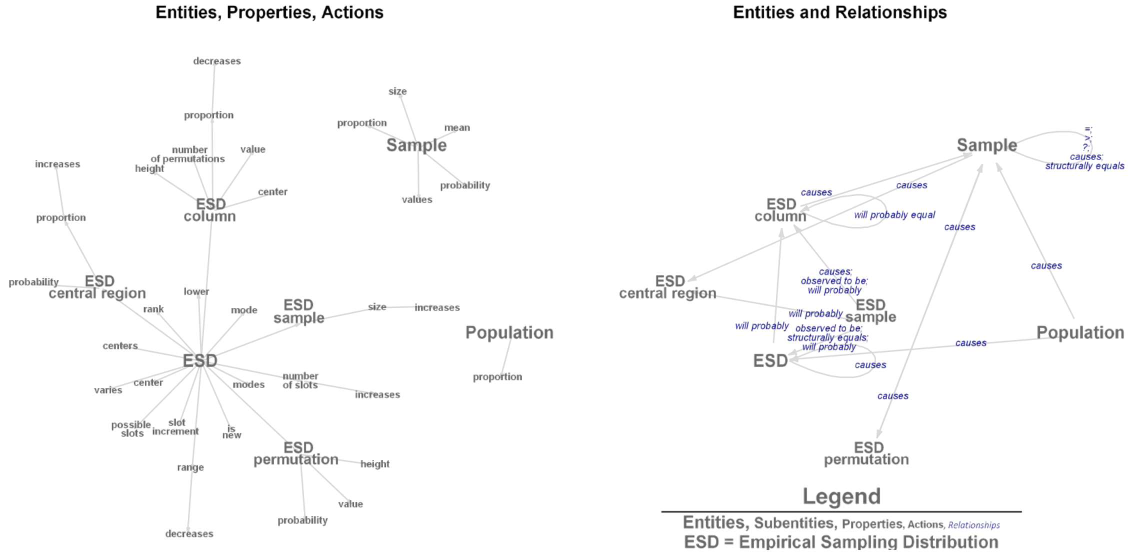


Figure 4.33. Summary of coded entities, properties and actions (left) and entities and relationships (right) for the Growing More Means: 0-1 activity for participant S.

S drew and described correct predictions for the ESDs with both sample sizes 2 and 3 (Figure 4.34). Their reasoning for predicting sample size 2 hinted at permutations already:

> Just because, like, if you—you could draw a zero and a one, or a one and—I mean, like a one and—you could draw a zero and a one, um, just it's more—I would say... Like it's—I would think it would be more likely dr—get, like—have an average of like point five then say like a one, because you would have to get two ones [OK], um, instead of like a 0 and a 1, or a 1 and 0. (3.1233–1242)

S listed the permutations that yield a mean of 0.5, and seemed to use those permutations for their assertion that 0.5 would be more likely, but the entities, properties, and relationships were ill-defined. S developed this reasoning slightly for their prediction of sample size 3, saying that they thought it would be easier to get "two of something, rather than three" (3.1300), and thought that means of 0.33 and 0.66 would be of equal height because there "shouldn't be any difference between, like, the chances of getting two zeroes and then also, like—or two ones" (3.1320–1322), since 0 and 1 are equally likely.

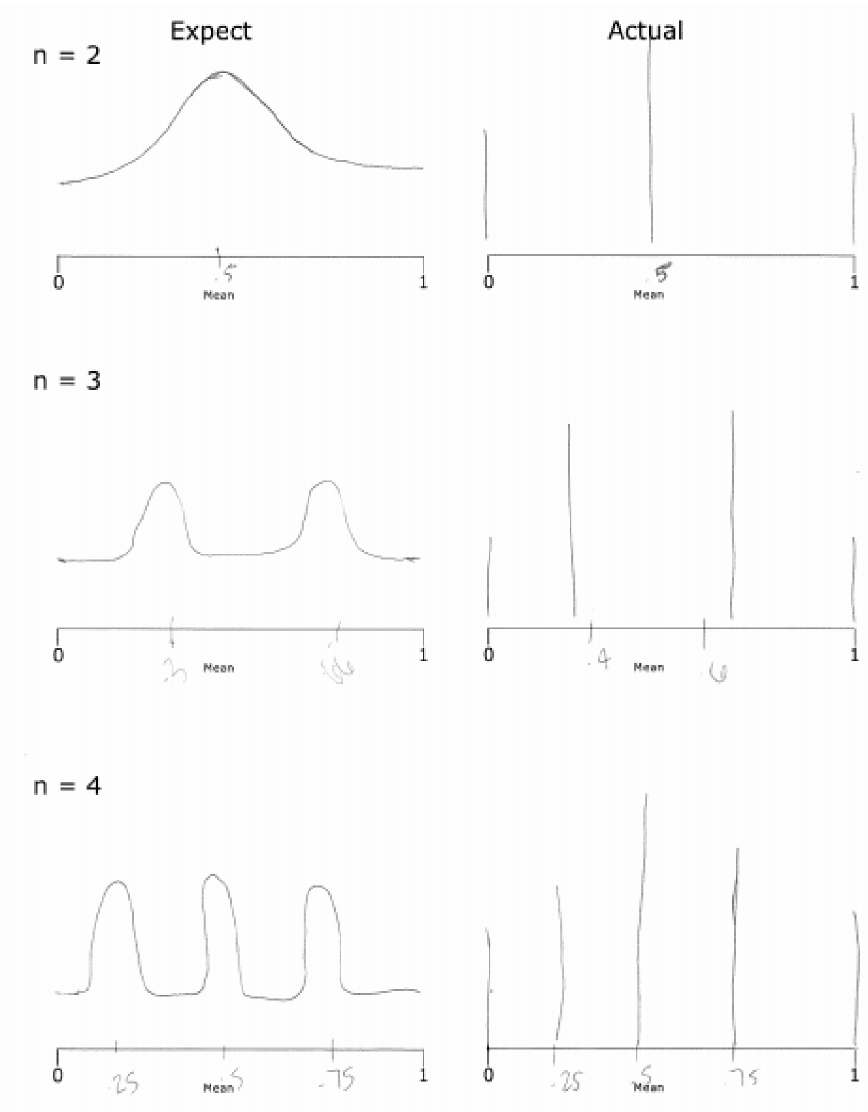


Figure 4.34. S's predicted (left) and observed (right) empirical sampling distributions for the mean of a 0-1 variable at sample sizes two (top), three (middle), and four (bottom). S drew the predicted, then observed the actual at each sample size before repeating the process at higher sample sizes.

However, S tried to apply this same reasoning to explain why the 0.25, 0.5, and 0.75 would all be equal: "You have the same chance of getting a zero or a one, um, for each time, it just depends on how many times you actually get them" (3.1356–1359). When confronted with an ESD that showed many more means at 0.5 than elsewhere, S explained that "it's a lot easier getting something in the middle" (3.1394–1395) because "it'll be less likely that you'll get, like, all four ones" (3.1399–1401). The implausibility of getting a mean of 1 or 0 seemed highly salient for S, and became a sort of shorthand for why means tended to heap in the center, which S drew on again when asked what they expected for the region between 0.4 and 0.6 if the sample size went up to 50:

> S: Um, I would say that the majority of the answers on the percent would increase for that range, um, just because you'll have less—less opportunity—I would say like it'll just kind of get—the range will get smaller.
>
> I: You were saying something about opportunity?
>
> S: Like, you're—it's gonna be a lot less likely that you're gonna get like zero per—like zero for the mean, because you would have to get like 50 zeros in a row, which is a lot more unlikely than getting like a couple ones and a couple zeros, so. (3.1480–1502)

S seemed to view this fact of the unlikeliness of getting 50 zeros as something that would lead the range to get smaller, but it was still not clear from their explanations why 50 zeroes would be more unlikely than any other mean. The unlikeliness of the 50 zeroes was a part of the mechanism that S was not able to articulate without further support.

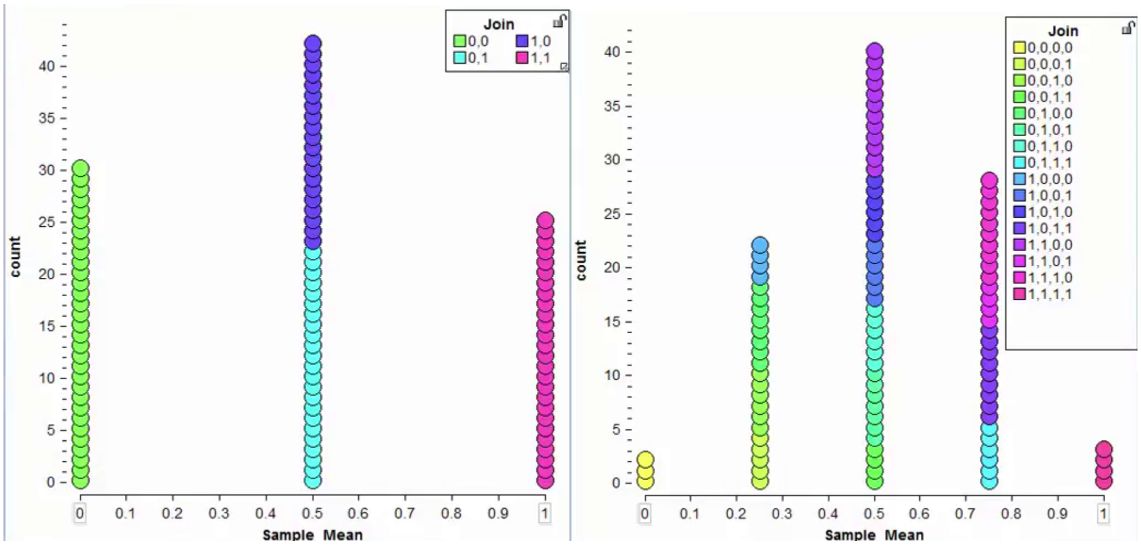


Figure 4.35. Empirical sampling distributions for a 0-1 variable at sample size 2 (left) and sample size 4 (right), colored and sorted by possible permutations for participant S.

S was able to go further when they viewed the ESD with the possible permutations indicated by different colors (Figure 4.35, left). When asked what they noticed upon viewing the plot, S noted “that it’s a lot more likely that you’ll get either a 0 1, 1 0, like, because that’ll equal the same mean, than you would for a 0 0, 1 1” (3.1553–1556). Both the color-coding and having a way to refer to the different permutations by the sequence as shown by the join attribute appeared to support S in reasoning here. Notice this reasoning was similar to their initial reasoning, but more specific with the addition of being able to refer to the specific permutation; S soon built on this and calls these entities “combinations”. The groups that included the permutations were now chunked in a way that S could parse the plot at sample size 4, despite its complexity (Figure 4.35, right):

> S: All of like, the—little stacks are kind of equal. The only one that's like much larger is, um, like the—the purple one for—so that would be like the 1 1 0 0, that's the most likely. At least in this trial it was. Um, but they all are about the same, like, number. [OK] So.
>
> I: And how do the number of combinations in each stack compare?
>
> S: They're about the same. They're just—it's just more likely to a mean of 0.5.
>
> I: And why is that?
>
> S: Because there's more combinations that you can do to get, um, a mean of 0.5. (3.1613–1633)

S notes, first of all that all the "little stacks"—the groups that are the same permutation—are about the same, so that all combinations are equally likely. S notices that one stack is bigger "in this trial", but seems to overall think that the stacks are essentially the same. When asked about the number of "combinations", S seems to interpret the question as asking about the number of sample means that are in each permutation stack, but goes on to clearly state that 0.5 is the most likely because there are more combinations that have that mean. From the hints of heaping reasoning S showed at the beginning, S is now able to name the key entities and properties of entities that operate in the mechanism of heaping.

Although the discrete nature of the variables made some aspects of the problem easier to track for S, and seemed to support their above reasoning about what exactly would be in the samples, it also presented challenges due to the cognitive energy that S expended on keeping track of how many possible means there were. Although S initially responded to a question about sample size increasing seeming to say that the region in the center would become less likely, it soon became clear that S was actually talking about the mean being *exactly* 0.5:

> Um, I think it would probably—it would get smaller, I think. [OK] They would still—I still think it would be more than like other numbers, but there's a lot—you're a lot more ch—like choices, or like you have a lot more areas that you—like there's more fractions available, like, for this one it's like—you only have a fourth of a chance, like 1 out of like—you can only really have like four—like 3—like I guess, like, I guess 5 different choices, but as you get like to like 8, you'll have 9 different things that you could have, um, [OK] 9 different probabilities, so. (3.1426–1439)

In addition to everything else, S was trying to keep track of the possible values, noting that the exact amount at the center was decreasing, and also noting that the region overall was becoming more likely. S repeatedly had to think about the what the possible options were, which may be easier for students who have higher numeracy. S later noted that as sample size increases, "it definitely—there's more variation, like—or there's not more variation, but there's more choices, with, like, smaller percents" (3.1465–1468). This was correct, but S had to struggle and at first apply the incorrect term "variation" when they were already confronted with a complex and unfamiliar situation. It is worth further study whether the unintuitive complexities of discrete variables are worth the tradeoff in trackability of other aspects of the mechanism (c.f. Wagner, 2006). S's sketches of predicted distributions for this activity all were curves, as would befit a continuous distribution, and they continued to draw curves after they had seen, and sketched, the actual results as discrete bars (Figure 4.34).

4.4.6 Growing More Means: 0-1-1

S's discomfort with non-equiprobable situations challenged their seemingly solid reasoning about heaping and empirical sampling distributions when they moved to a 0-1-1 variable. Even though there were several scaffolds in place to help them reason about the shape using permutations, they struggled to do so, probably because there was many more permutations to track, but also because they did not seem to have precise understanding about the population process and what the expected mean would be. This was consistent with their earlier responses in Growing Sample Mean: 0-1-1Growing Sample Means: 0-1. Therefore, there were much fewer codes regarding permutations, and there was much less precision in the identification of entities and their properties (Figure 4.36, left). Note that S never was identified the *Population mean* or even specified the probability of drawing a 1, instead resorting to words like “higher” or “more”. S made many guesses about the ranking of the heights of columns in the ESD, and attempted to draw causal connections between the different entities (Figure 4.36, right), but these connections were not generally grounded in any sense of permutations.

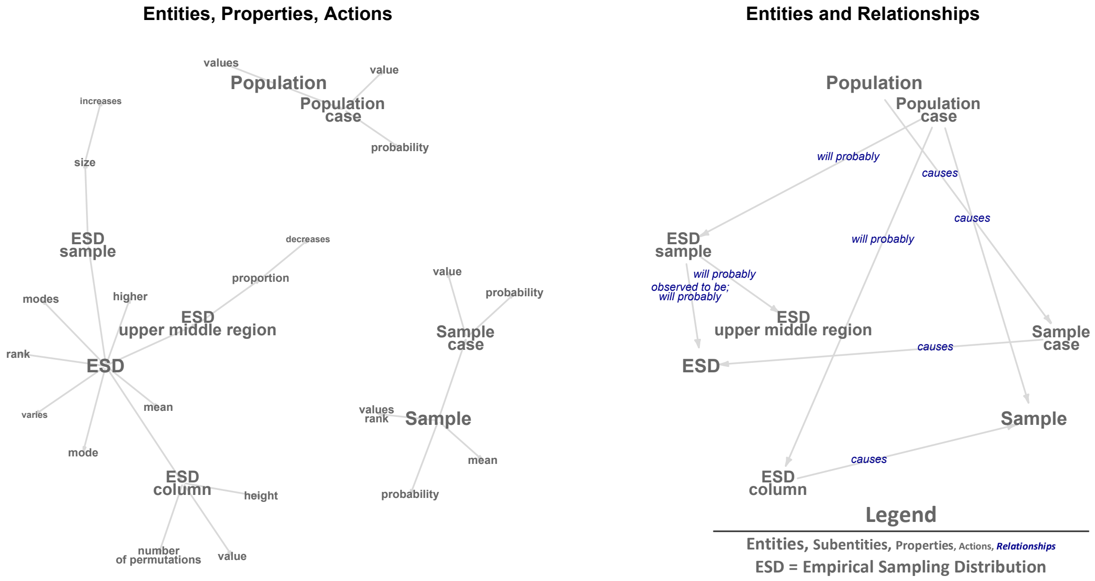


Figure 4.36. Summary of coded entities, properties and actions (left) and entities and relationships (right) for the Growing More Means: 0-1-1 activity for participant S.

S both struggled to predict and explain the behavior of the ESD in the 0-1-1 situation (Figure 4.37). They knew that the mean should move “more towards, um, one just because you have another one in the mix” (3.1529–1531), but their first predicted plot did not even match the possible values for sample size 2, and they were surprised that 0.5 was so high given the increased likelihood of 1s. Even after seeing this, they still predicted 1 would be highest at sample size 3, because “you have three times that you’re taking it, and you’re gonna be more likely to get a one than a zero” (3.1714–1716). Immediately after reasoning successfully with combinations and about possible values for samples for the 0-1 case, S’s reasoning was almost entirely driven by a vague sense of the increased probability of ones, leading them to predict everything piling up at a mean of 1 even after seeing some slightly disconfirming evidence of this at sample size 2.

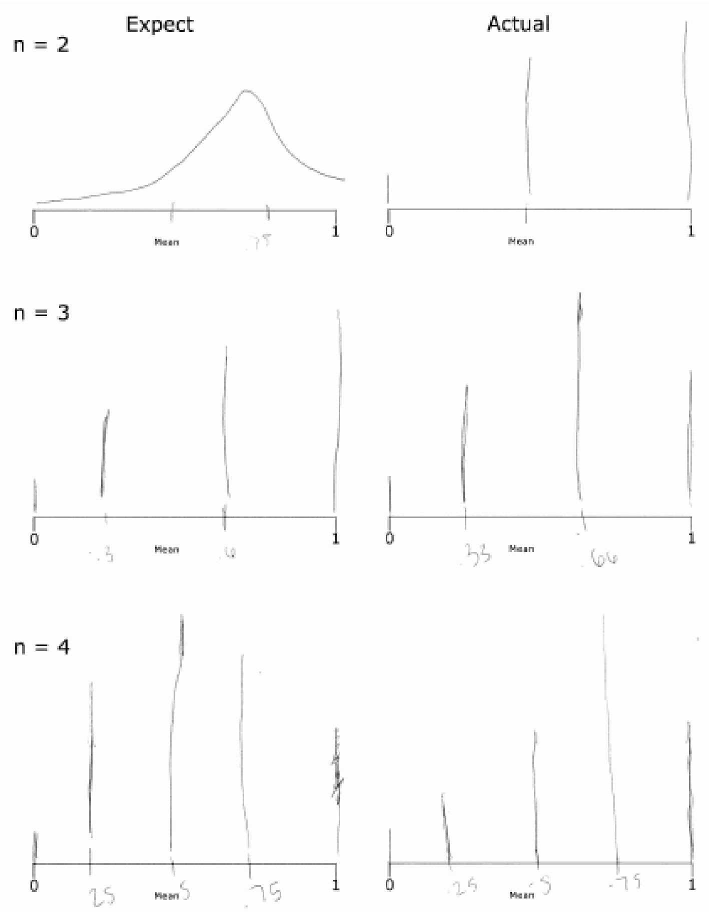
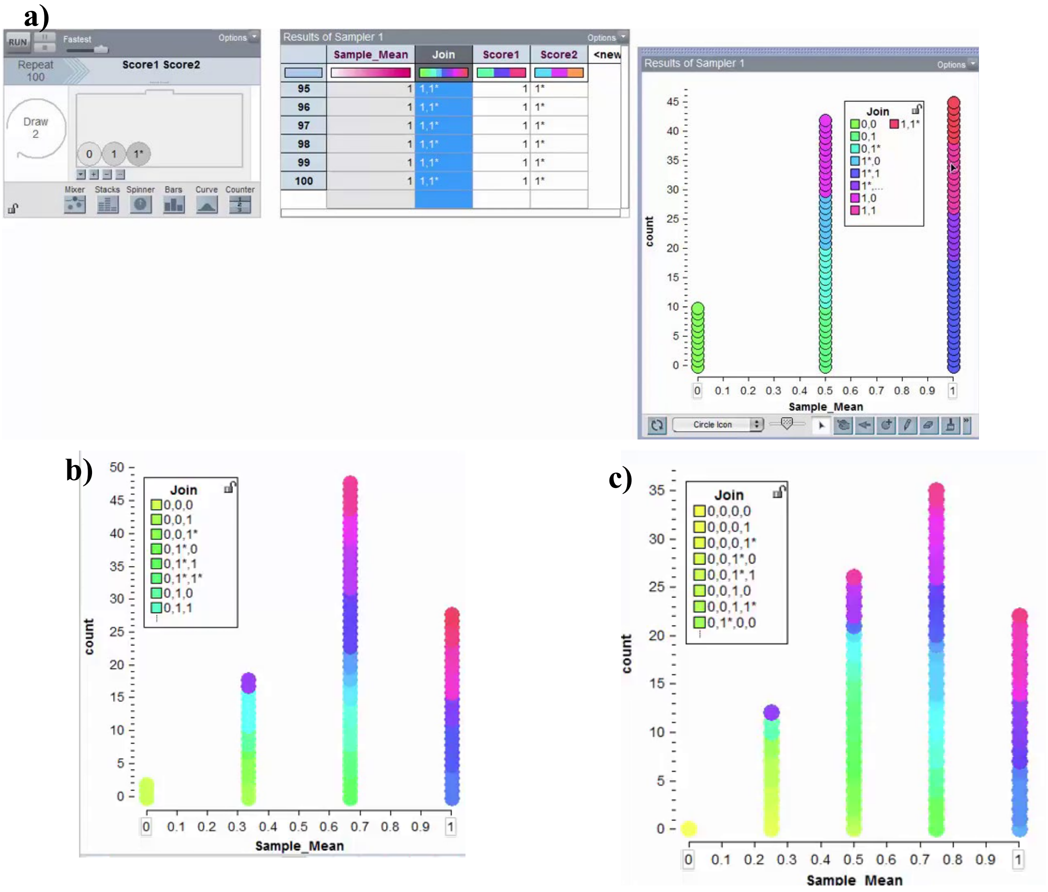


Figure 4.37. S's predicted (left) and observed (right) empirical sampling distributions for the mean of a 0-1-1 variable at sample sizes two (top), three (middle), and four (bottom). S drew the predicted, then observed the actual at each sample size before repeating the process at higher sample sizes.

Once S saw the results for sample size 3 (Figure 4.38b), their reasoning started to shift as they saw that some other factor must be at play besides means of 1 becoming more likely, again drawing on the term "combination":

> Um, cuz it's more—there's more ones in the mix so while—before there's probably like an equally likely chance, like, you're gonna be more likely to have the combination of having, like—maybe like two ones and a 0 rather than like all ones, or like, all, like, basically all zeros and then like one one. [OK] So. (3.1741–1749)

Whereas S previously used the term "combination" to denote different possible sample permutations, which allowed them to reason effectively about why certain values were more likely, here it appeared that S was actually referring to combinations, i.e. the number of zeros and 1s which determine the possible means. S remembered that 0.5 had the most combinations in the 0-1 case, and appeared to be using this as evidence that there were two opposing forces: 1) the "more ones in the mix" drawing the mean upward to one, and 2) more combinations at 0.5 which meant that "two ones and a zero" would be more likely than "all ones". S compensated for their previous error of overestimating the percentage of sample means at 1 by estimating that 0.5 would be highest at sample size four (see Figure 4.37), explaining that "you're gonna have a lot more combinations getting the 0.5, um, so that's where I think 0.5 would be the most" (3.1771–1775). Seeing that this was not in fact true (Figure 4.38c), S created a new rule and predicted as sample size grows that "whatever fraction is, like, closer to one is always going to be, like, the top" (3.1796–1798). This expectation was not accompanied by the clear description of the mechanism that would lead to that result, although it did describe the observed graphs at sample sizes 2, 3, and 4. S's *ad hoc* explanations did have hints of mechanistic reasoning, since both the population percentage of 1s “in the mix” and the permutations/combinations were mechnistically relevant, but S did not achieve integration or clear identification of entities, properties, or activities.



Results of Sampler 1

	Sample_Mean	Join	Score1	Score2	<new
95	1	1,1*	1	1*	
96	1	1,1*	1	1*	
97	1	1,1*	1	1*	
98	1	1,1*	1	1*	
99	1	1,1*	1	1*	
100	1	1,1*	1	1*	

Figure 4.38. TinkerPlots™ setup and permutation plots for a) sample size 2, b) sample size 3, and c) sample size 4 for participant S. Both 1 and 1* were scored as 1, with 1* indicated separately to distinguish equally probable permutations.

In addition to S's more vague sense of the population process, several differences in the way this task was administered may partially account for why S did not attend as much to permutations as they did in the previous activity. In the 0-1 version of this task, S actually went through the entire cycle of prediction and viewing, and attempting to explain those results, *before* then viewing the permutations and being asked pointed questions about what the permutation plots told them. In the present activity, the permutation plots were showing the entire time, and S was not asked any questions to draw their attention to those plots. Another complication was that the meaning of having the outcome "1*" in the sampler was intended to separate equally likely permutations in the plot, but this intent was never explained. S never was asked to examine the combinations, to see how the combinations involving 1* differed from the combinations involving 1, never referred to the 1* outcome, and simply said phrases like "there's more ones in the mix" (3.1742–1743). Without support for any meaning for the 1* designation, S seemed to ignore it. Finally, as Figure 4.38 shows, the plots themselves were much messier because there were so many possible permutations. This both had the effect that the permutations were hard to distinguish perceptually using the default colorings in TinkerPlots™, and also that the heights of the combinations were much noisier since there were only 100 simulated means total. Therefore, some of S's lack of attention to permutation and to the mechanism can likely be attributed to these drawbacks in the task and prompt design.

4.5 Growing Possibilities & Many Means

Although this interview introduced theoretical sampling distributions (TSD), S seemed to make little use of the entity outside of the two sections where S was explicitly manipulating TSD permutations (Table 4.5). S mainly used TSDs as a tool for thinking about the likelihood of sample means. S had difficulty seeing connections between samples that had different sample sizes across different TSDs, and so generally there was more reasoning about the movement of sample means as sample size increased when they were examining a sample size plot.

Table 4.5
Presence of Entity, Property, and Action Codes that Occurred in More than One Section of Interview 4 for Participant S

Entity	Property	Action	Presence in Each Interview Section					
			GP 0-1-1 Blocks	GP 0-1-1 SSP	GP 0-1-1 Blocks	GP 0-1-1 SSP	G++M 0-1	G++M 0-1-1
ESD			_	_	_	_	■	■
ESD	number of slots		_	_	_	_	■	■
ESD	number of slots	increases	_	_	_	_	■	■
ESD sample			_	_	_	_	■	■
ESD sample	size		_	_	_	_	■	■
ESD sample	size	increases	_	_	_	_	■	■
Population			■	_	■	■	■	■
Population	values		_	_	■	_	_	■
Sample			■	■	■	■	■	■
Sample	mean		■	■	■	■	■	■
Sample	mean	centers	_	_	_	■	■	■
Sample	mean	changes	_	■	_	_	■	■
Sample	mean	decreases	_	■	_	_	■	■

Sample	mean	increases	___■■_
Sample	mean	stays	_■__■■
Sample	mean	zigzags	_■_■■■
Sample	number of slots		_■__■_
Sample	probability		■■■___
Sample	probability	decreases	_■■___
Sample	size		_■■■■■
Sample	size	increases	_■_■■■
Sample case			_■__■■
Sample case		is added	_■__■_
TSD			■■■___
TSD	number ending white		■_■___
TSD	number of permutations		■■____
TSD	number of permutations	increases	■■____
TSD	number of slots		■_■___
TSD	number of slots	increases	■_■___
TSD column			■_■___
TSD column	height		■_■___
TSD column	number of permutations		■_■___
TSD column	value		■_■___
TSD Sample			■■■___
TSD Sample	size		■■■___
TSD Sample	size	increases	■■■___

Note. A symbol of _ indicated that the code did not occur within that interview section, whereas ■ indicated that the interview section contained the code. G++M = Growing More Means, GP = Growing Possibilities, SSP = Sample Size Plot.

4.5.1 Growing Possibilities: 0-1 Building Blocks

S spent most of their time in the building blocks activities figuring out the possible permutations for each sample size and answering narrowly phrased questions that were intended to support them in seeing connections between different sample sizes. These responses were not coded mechanistically. However, there were some conceptual questions here, and many of the responses concerned a new top-level entity, the TSD, which here was represented by building blocks (Figure 4.39), and which unsurprisingly was dominant in the mechanistic codes (Figure 4.40, left). Because of the way the prompts were asked, some coding revolved around the number of permutations ending in a white block (*TSD number ending white*) or in a black block (*TSD number ending black*). Each element of the TSD was a possible sample permutation, and hence there were properties associated with *TSD Sample* as an entity, as S often compared properties of a possible sample mean (*TSD column*) and the permutations within it. As in the Growing More Means activities, certain responses could have been coded either as an element of the TSD (*TSD Sample)* or as a sample not necessarily linked to the TSD *(Sample)*, and these codes are likely interchangeable. By the end of this segment, S was expressing a rich network of causal links across nearly all the entities, especially in relating the number of possible samples in a TSD column to the likelihood of a sample mean being in that column (Figure 4.40, right).

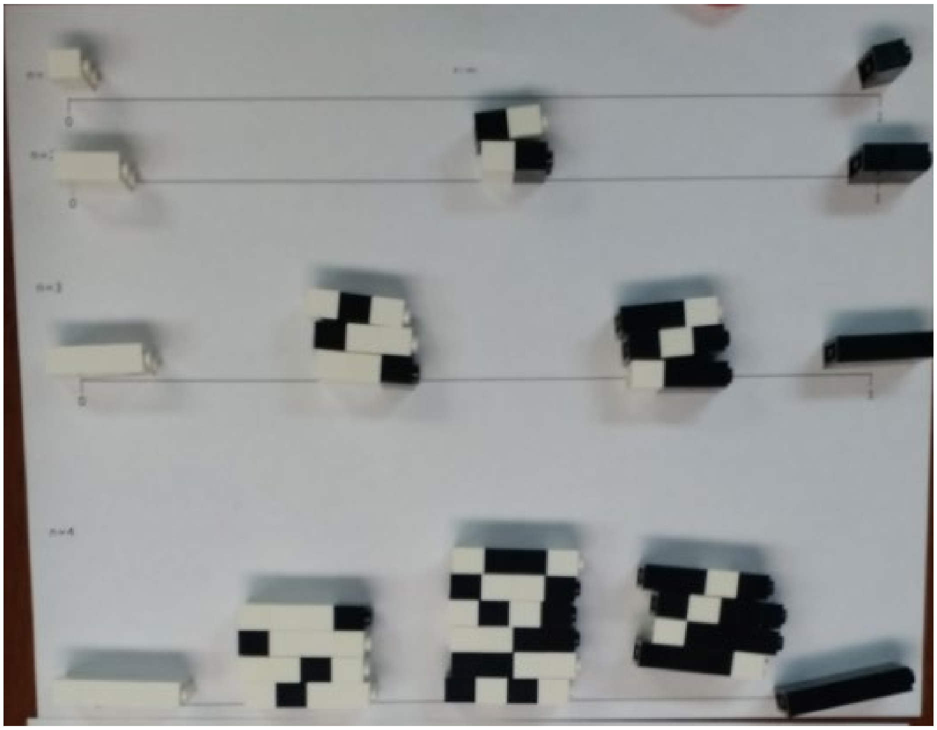


Figure 4.39. Theoretical permutations plots for 0-1 situation created by S using building blocks for sample sizes 1 (top) through 4 (bottom). Black represents a value of one and white represents a value of 0; each set of building blocks stuck together represents a possible sample whose mean is placed on the x-axis at its appropriate mean.

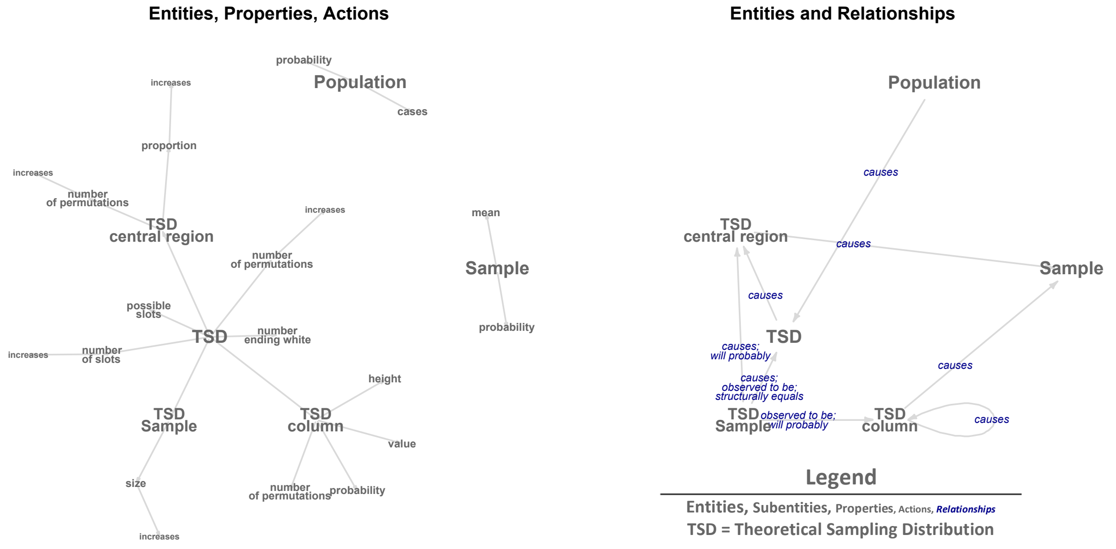


Figure 4.40. Summary of coded entities, properties and actions (left) and entities and relationships (right) for the Growing Possibilities: 0-1 Building Blocks activity for participant S.

Reasoning about possible sample permutations as first-class entities helped S build on the mechanistic reasoning that they had displayed in Growing More Means: 0-1. At sample size 2, S noticed that the permutations were increasing, and explained that 0.5 was the most likely because “there’s two different combinations that you could get to get 0.5, while, um, to get 0 or 1 there’s only one combination that you can get” (4.164–167), their most precise statement of heaping yet so far. S started to grapple more with the relationship between permutations and the means, noting at sample size 3 that there could be different permutations with the same mean in the center: “it’ll be the same number for that mean, but it’ll affect, like, what order it is, and it’ll become—it’ll make it more probable that it could happen” (4.258–263), a chain of reasoning back and forth in the mechanism where

S connects the possibility of multiple orders in a column to the higher probability. A novel feature of S's reasoning in this session was that their attention was drawn to why there were more possibilities at the *center* of the distribution, whereas in Growing More Means: 0-1, S primarily focused on how the extremes of 0 and 1 would be more unlikely. Whereas S previously had explained the increasing concentration of permutations in the center at high sample sizes by noting that drawing 50 zeroes would be so unlikely, S here was able to explain in the inverse manner:

> I: So if we went up to N equals 10, would you think that zero would be more likely here than at N equals four, less likely, or about the same?
>
> S: Um, less likely.
>
> I: Okay. And why is that?
>
> S: Cuz there's more combinations an—there'll be more combinations in the middle that you could get that aren't zero, [OK] so it just put it down even lower for percentage-wise. (4.420–435)

S now explained how *zero* is more unlikely because it has the same number of combinations, but the combinations are rapidly proliferating in the center, and so the zero combination becomes proportionally a smaller part of the whole. This ability to move around different parts of the mechanism, and chain both directions, appears to represent an advance in their mechanistic reasoning about heaping.

However, the quirks of discrete distributions still played a role in S's reasoning. When asked about what would happen in the middle region at sample size 50, S drew again on thinking about how many possible slots there are:

> Um, I think... that... they'll even—it'll be, um—like most of them will also still be in that range. I think it'll be a higher percentage [OK] than t—n equals 10, just because the

> fractions are getting smaller, um. [OK] So, it'll just like put them closer together and t—closer or they'll be more—like different of outcomes, like an increase in—and then in that, um, range. (455–465)

The grouping in the center was not based primarily on permutations, but on the fact that "the fractions are getting smaller", making them "closer together". Given S's usage of the word "fraction" extensively in the Growing More Means activities to refer to the number of possible slots or places that have possible means, this appeared to be part of the foundation of their reasoning for heaping. Although the increase in slots was in fact true, it did not particularly explain why the central range would have more values. *All* ranges, including the extreme ranges, would also have more possible values.

An apparent downside of this activity was the cognitive energy that S devoted to thinking of all the possible permutations. It was intended that participants would learn to quickly build up permutations for the next level by simply adding a white block, and black block, successively to all the permutations at the current level, and several supports and scaffolds were put in place to encourage this reasoning. Nevertheless, S did not use that strategy and struggled to think of all the possible combinations. At some times, this struggle did appear to relate to core aspects of the mechanism, such as when S realized, "I have to add one more on this side, too […] just because the proportions are the same" (4.231–233), noting that because of the equal probability of white and black blocks that the distribution should be symmetric. However, at other times, their struggle seemed more in the realm of extrinsic than intrinsic cognitive load, as S struggled to find the permutations for sample size 4:

> Yeah. I'm kind of looking at the ones before, and then I'm also thinking of, like, if I take one block away, then I could make it, like, black or white, or just, like—w—like—just like switching them around and then also like thinking what could I start with that still ends with like a white block, [OK] or what could end with that still ends with a—a white block. (4.272–280)

S initially failed to find all permutations and thus had equal piles at 0.25, 0.5, and 0.75. It is not clear how much S's struggles facilitated learning about the mechanism of sample size. The mechanism of sample size was hypothesized to be more helpfully illustrated by the connections between levels rather than S's struggle to find the missing permutations.

4.5.2 Growing Possibilities: 0-1 Sample Size Plots

After building a theoretical sampling distribution out of permutations, S was asked to visualize several of the possible samples in the distribution as sample size plots. For instance, a sample consisting of the permutation 0-1-1-1 would be visualized as a sample size plot with the mean starting at 0 at sample size 1, and then moving to 0.5, 0.66, and finally 0.75. S then was asked to draw what they expected the sample size plot to look like if more blocks were randomly added to that sample as the sample size increased to 25. Although graphing the sample size plots of the permutations was intended to help draw connections between the swamping and heaping, S seemed to regard these two representations as disconnected entities and much of this segment was spent clarifying the representations and questions rather than attempting to describe anything mechanistic. The existing entities and properties were generally a subset of prior swamping codes with a sprinkling of codes concerning the TSD (Figure 4.41). S entirely reasoned using swamping reasoning based on the sample size plot for most of the activity, and abruptly switched to heaping reasoning based only on the TSD at the end of the activity, with no connections expressed between the two levels (Figure 4.41, right).

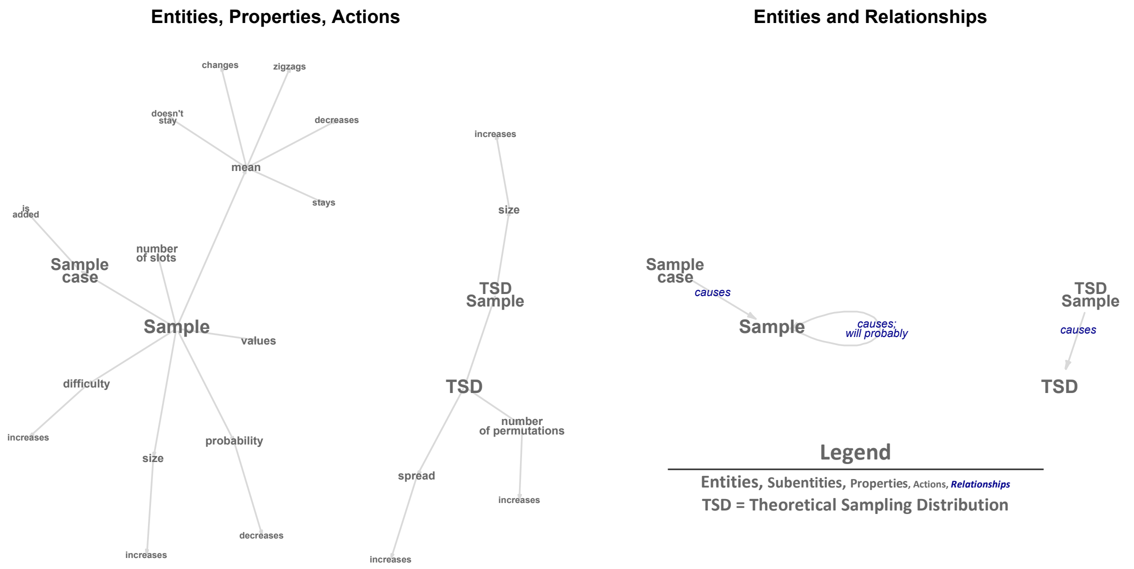


Figure 4.41. Summary of coded entities, properties and actions (left) and entities and relationships (right) for the Growing Possibilities: 0-1 Sample Size Plots activity for participant S.

S had difficulty translating samples represented as building blocks onto the sample size plots, despite their apparent fluency with both representations separately. S struggled to see how a sample of all black blocks would be represented on the sample size plot. Once they had made this translation, S then had difficulty transitioning to drawing a sample size plot of what they expected to happen as the sample size grew from 4 to 25 (Figure 4.42). Some of this may have been due to the unfamiliarity of the task, the unclear instructions, and the inherent challenge of switching between very different representations. However, even once S did apparently understand the task, there was an intriguing trace of heaping-like reasoning in their response:

> Well, each black or each white will definitely affect it, and it—it might go, like, back—it'll probably go back and forth, um, each time, cuz, like, each time you'll probably get a mixture of black and white (4.725–730).

S refers to the impact of each additional case on the mean at first, and refers to zigzagging, both similar to their earlier swamping reasoning. However, the reason for that zigzagging is now because "each time you'll probably get a mixture of black and white". This phrase could indicate just that they could get either one on an individual draw. Yet it would be a strange way of saying what would happen "each time", since each time only one thing will happen. Another possibility is that S was chunking groups of additional draws into "each time", and seeing that if adding group of more draws, that the mixtures of black and white are most likely because of heaping. A third possibility is that S is referring to the entire sample of 25 values as "each time", but this also does not entirely align with S's statement, because that does not explain why the line would go "back and forth". These data do not provide enough evidence to distinguish these three possibilities, but this was the first time S has referred to the possibility of getting a "mixture" in terms of the sample size plot, a term that S would come back to in the post-interview. After a bit more practice, S settled into a routine of predictions similar to "it'll probably zigzag around for a little bit, um, but after quite a few trials, it—or, the sample size increases, it should be about a 50/50" (4.814–

817), marking the return of both “zigzag” and “50/50”, neither of which S had mentioned so far in this interview.

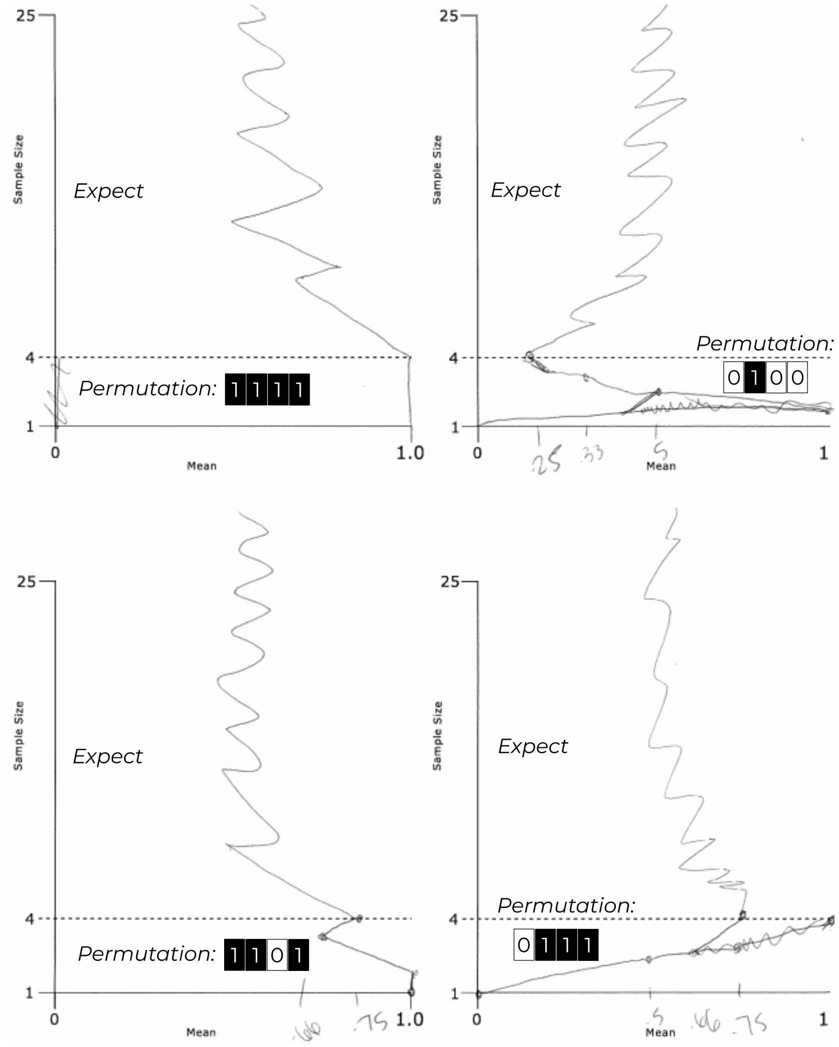


Figure 4.42. S's sample size plots based on permutations of building blocks for a 0-1 situation. The interviewer pointed to a permutation of four building blocks, representing a possible sample of size 4; S drew the sample size plot representing the movement of the mean from sample size 1 to 4, and then drew what they expected to happen if the sample size continued to grow to 25. Italicized text and block visualization were added later for clarity.

S’s final response at the conclusion of drawing these sample size plots completely ignored the sample size plots and went back to discussing permutations, perhaps cued by the word “distribution”:

> I: And so, um, what do you think will happen—and so, what do you see happening to the distributions as the sample size increases?
>
> S: Um, they get more spread out, but there's just more of each, under—like for—there—cuz there's more—there's different, like, combinations that you can get, and there's more of them for each time you increase the sample size. So it'll just keep on increasing, um, the number that you could potential—like the—the number of potential combinations that you could get each time you draw [OK] four blocks—or however many sample size there is. (4.841–857)

It is unclear what S means by “more spread out”—earlier, they had referred to the increase in number of slots as more “variable”. The feature that stood out to them among all the features the noticed previously appears to be that there are more permutations as the sample size grows. Although this is true, it is not very focal to the mechanisms of sampling variability; more central would be the proportional increase in permutations in the center that S had previously noted.

4.5.3 Growing Possibilities: 0-1-1 Building Blocks

Similar to S’s previous experiences with 0-1-1 populations, S lacked a clear sense of the long-run center and simply noted that the mean would be higher. Perhaps because S did not have as many accessible intuitions and resources around the process, their mechanistic reasoning was somewhat less rich (Figure 4.43, left). Mechanistic codes focused on answering questions about the likelihood of various sample means (*Sample mean probability*) and noticing corresponding changes in the heights of the columns of the theoretical sampling distribution (*TSD column*). S again never identified the population mean, and did not directly causally link the population setup with the TSD (Figure 4.43, right).

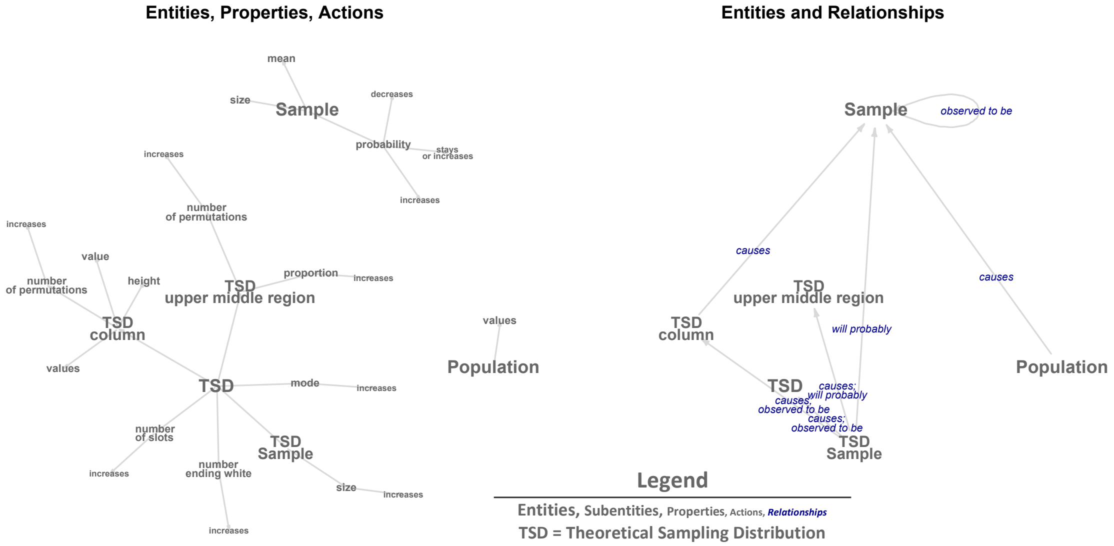


Figure 4.43. Summary of coded entities, properties and actions (left) and entities and relationships (right) for the Growing Possibilities: 0-1-1 Building Blocks activity for participant S.

S was able to create the theoretical permutations for the 0-1-1 situation relatively easily (Figure 4.44), but again never expressed the population mean or the expectation that would have for the theoretical sampling distribution, only mentioning that "you're more likely to be able to get a higher mean rather than a lower mean" (4.964–967), and that as the sample size increased, that the "combination where it's like the one that's closest, or m—it's gonna be more right, on, like, the right side of the graph" (4.976–980). The "one that's closest" may be referring to their hypothesis in Growing More Means: 0-1-1 that the column that was closest to 1 would be the tallest. Since S had not yet seen sampling distributions for 0-1-1 with any sample sizes higher than 4, they had not seen any disconfirming evidence to this hypothesis. It seemed that without the symmetry of the 0-1 situation to guide them, they again had trouble arriving at a clear sense of what would happen as the sample size increased. This inability to recognize the phenomenon was both a reflection of their lack of mechanistic understanding—the connection between the population average and the expected average of the sampling distributions—but also probably hampered developing their mechanistic reasoning further since they did not have a sense even of what the mechanism would produce.

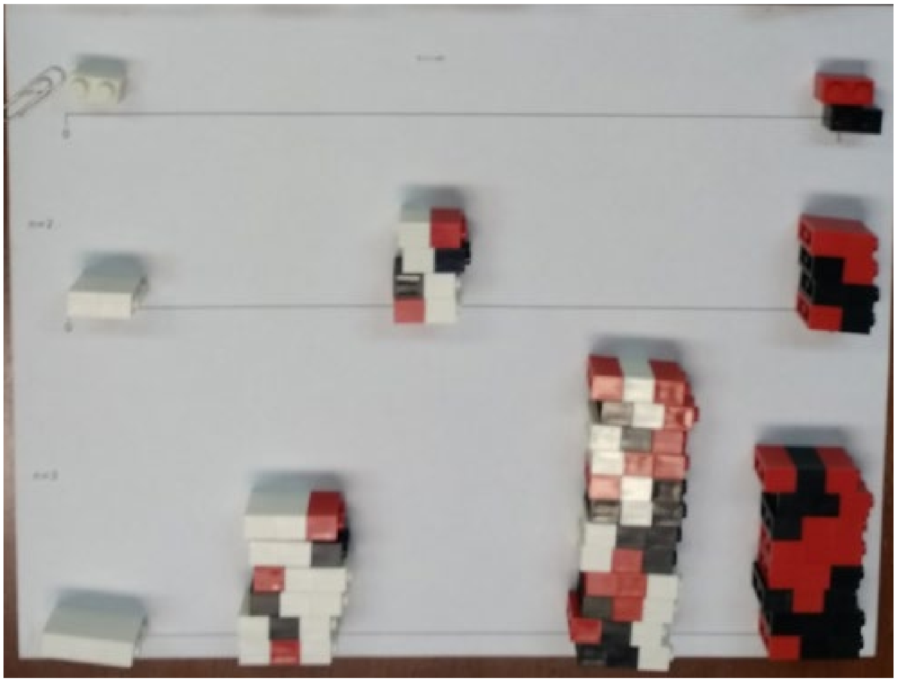


Figure 4.44. S’s theoretical permutations plots for 0-1-1 situation created using building blocks for sample sizes 1 (top) through 3 (bottom). Black and red represent values of one and white represents a value of 0; each set of building blocks stuck together represents a possible sample whose mean is placed on the x-axis at its appropriate mean.

S’s responses revealed a difficulty in tracking heaping through permutation plots, which was that noticing the percentage increase in the center required comparing *relative* rates of change. The fact that some possible means had more permutations every time, and yet still had a percentage decrease, was more difficult for S to coordinate than to see that 0, which stayed at just one permutation, would experience a percentage decrease:

> I: OK. And, um, at say N equals 10, would zero be more or less likely than here at, um, N equals fou—n equals three?
>
> S: Um, less likely just because the—uh, like the other probabilities are just increasing while that one's just staying the same, so.

> I: Okay. And what about at one? Would one be more likely or less likely?
>
> S: Um, I'd say it would get more and more likely. But, {yeah}, yeah, because it—the—the combinations are only going to increase. Um, so I would say it's probably gonna become more [OK]—more likely or stay around, like, the same, like, area. So. (4.982–1002)

S realized that 0 would decrease in percentage, just as they were able to in the 0-1 situation, because it was constant and everything else decreased. However, their initial response was that 1 would also get more likely, because "the combinations are only going to increase", although they soon qualified that it may also stay around "the same, like, area". Although the permutations at 1 will double every time, the permutations near 0.66 will grow even faster, a complex thing to imagine in an already abstract scenario. S was never asked to compare the relative frequency for a mean of 1 as the sample size increased; if they had, they might have noted that while the absolute frequency doubled at each increase in sample size, the relative frequency will decline since the size of sample space tripled.

4.5.4 Growing Possibilities: 0-1-1 Sample Size Plots

S now drew sample size plots starting with several of the possible samples they had created for the 0-1-1 scenario, analogous to Growing Possibilities: 0-1 Sample Size Plots. Due to time constraints, S was asked rather less probing and "why" questions during this part of the interview, which is probably part of why the codes were relatively minimal and there was no mention of the theoretical sampling distribution (Figure 4.45).

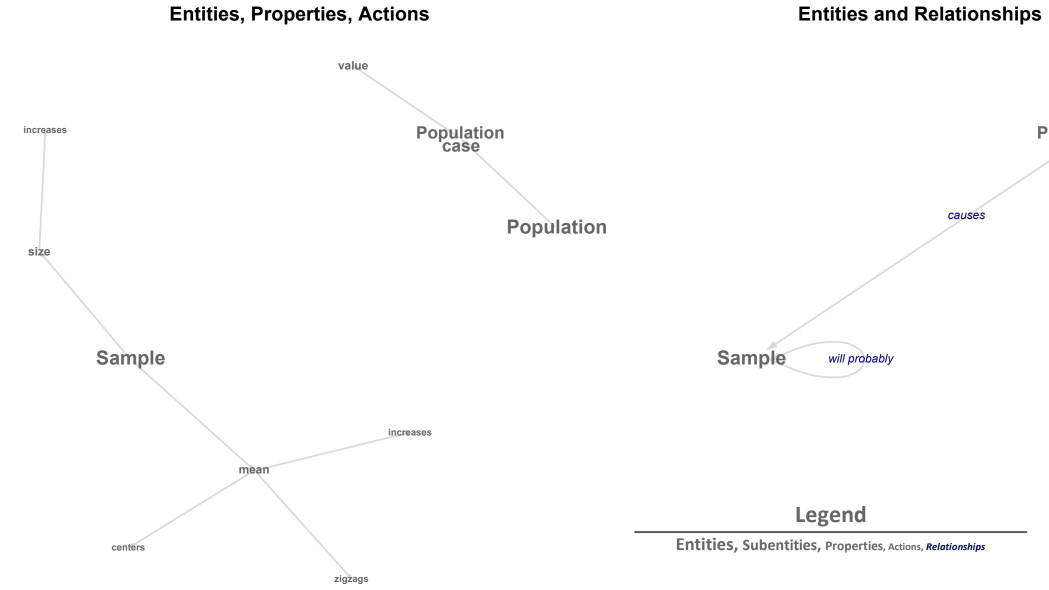


Figure 4.45. Summary of coded entities, properties and actions (left) and entities and relationships (right) for the Growing Possibilities: 0-1-1 Sample Size Plots activity for participant S.

S drew relatively similar sample size plots for all four samples chosen by the interviewer (Figure 4.46). However, the first drawn (Figure 4.46, upper left) had relatively similar size of the zigzags as sample size increased, and S spoke about it going "back and forth". The interviewer asked a question that seemed to subtly change S's responses thereafter:

> I: Okay. And would it keep—would the—um, would these back and forth zig zags stay about the same, or would they...
>
> S: It would get smaller as you go up I would say, probably. (4.1047–1054)

When asked about the size of the zigzags, S reiterated that the zigzags would decrease as sample size increased, which they had stated before for the 0-1 sample size plots. However, the remaining sample size plots had noticeably smaller zigzags as they got closer to 25, and S repeated this assertion for the remaining sample size plots. Additionally, S used the word "zigzag" in their remaining responses; although S was the one to introduce this word during the Growing Sample Proportions section, S had not used the word so far in this section until the interviewer reintroduced it. This incident serves as a reminder that students may only say what they are given opening and opportunity to say, and what they are in various ways cued to say. As a corollary, the sparsity of mechanistic reasoning in this section is likely due to S being given fewer opportunities to reason mechanistically.

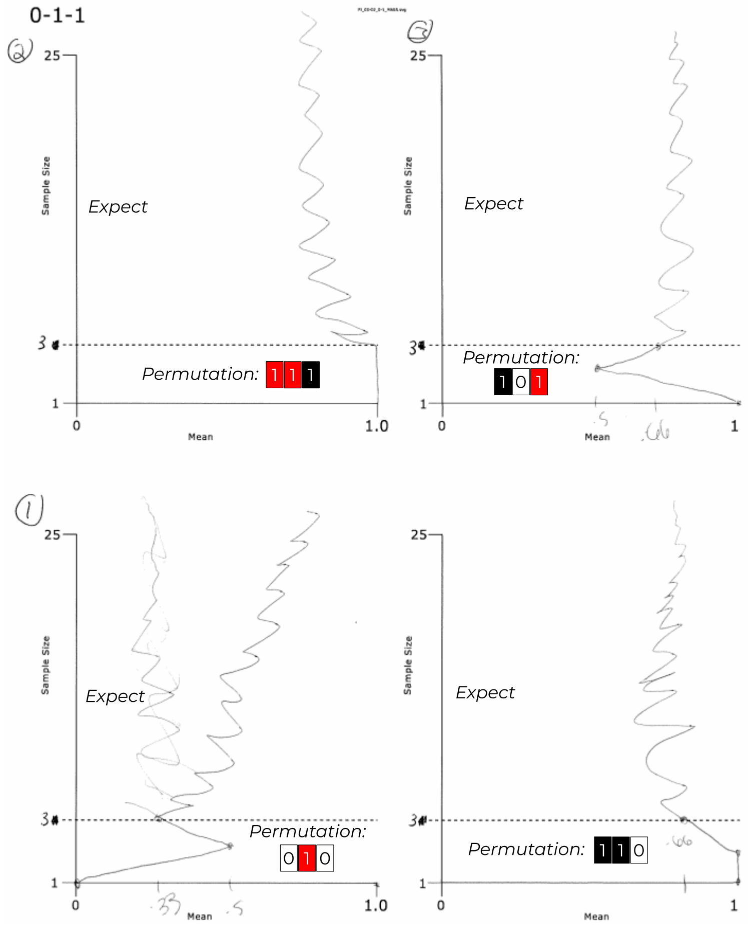


Figure 4.46. Sample size plots based on permutations of building blocks for a 0-1-1 situation. The interviewer pointed to a permutation of four building blocks, representing a possible sample of size 3; S drew the sample size plot representing the movement of the mean from sample size 1 to 3, and then drew what they expected to happen if the sample size continued to grow to 25. Italicized text and block visualization were added later for clarity.

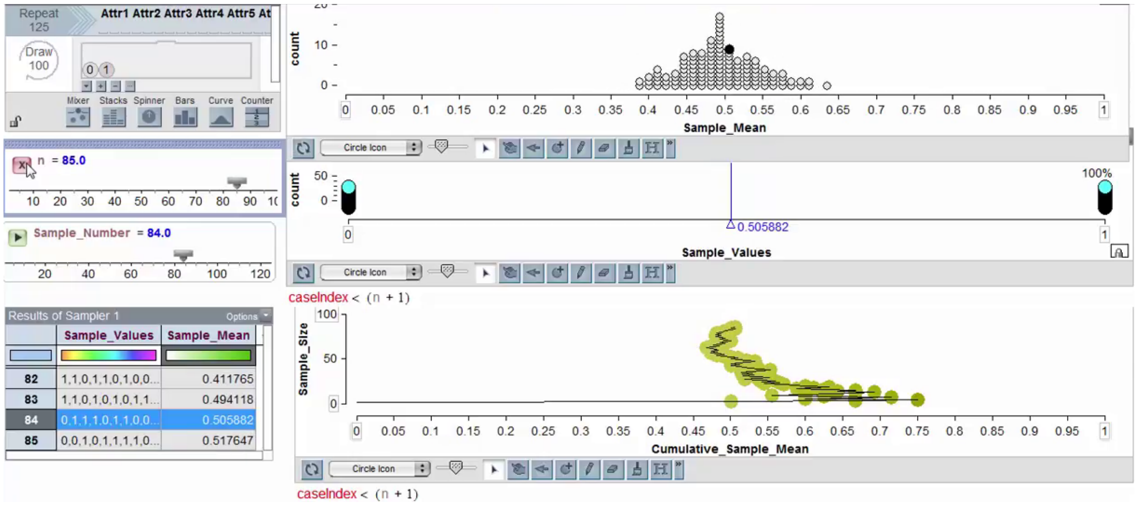
However, S had an opportunity to show more mechanistic reasoning in their third plot, where they initially drew it centering on the left side (Figure 4.46, lower left):

> So the white, red, white. So it would start at one, um, and then it would go to a point five, and then it would go to point three three, and then—for it going up to like 25, you'll probably go back and forth for a while and kind of get s—like, less zigzaggy as you go up. [OK] And probably centered somewhere on the left side of, um, the graph. [OK. And...] Or wait. No, that's wrong. [OK] Okay, it would go more towards—it'll probably end up, um, going more towards like the right, and kind of zigzag more to the right. Just because there's more of like the green—or not green—the red or black blocks that would definitely shift it over [OK] to one. (4.1091–1105)

This block sequence was the only one of the four samples to have a mean on the left side of the graph, and S's original inclination was to have the mean centered on the left. They then realized that they would have a higher chance of getting ones than 0s, but again without a specific sense of center—only as specific as "shift it over to one".

4.5.5 Growing Many Means: 0-1

After several activities focusing on the sampling distribution, and drawing some connections with the sample size plot, S then interacted with a complex interface involving linked representations of the sample size plot, the sample, and the empirical sampling distribution (Figure 4.47). This representation led to many rich descriptions of the movements of the mean of a sample, with no reference to the TSD (Figure 4.48, left). Revealingly, S's mechanistic reasoning was primarily about the movement of the means and the sample size plot, and they paid little attention to the properties of the empirical sampling distribution beyond the fact that the number of "slots", or possible sample means, increased with the sample size (Figure 4.48, right). This 0-1 situation was back in S's comfort zone and so they again were able to note causal connecitons between the population proportion of 1s and what happened in a given sample.



n = 85.0

Sample_Number = 84.0

Results of Sampler 1 Options

	Sample_Values	Sample_Mean
82	1,1,0,1,1,0,1,0,0...	0.411765
83	1,1,0,1,0,1,0,1,1...	0.494118
84	0,1,1,1,0,1,1,0,0...	0.505882
85	0,0,1,0,1,1,1,1,0...	0.517647

Figure 4.47. TinkerPlots™ setup and results for Growing Many Means: 0-1 for participant S. Upper left shows 0 and 1, equally likely; middle left has sliders for controlling sample size and which sample is highlighted on the top right and shown in the middle and lower right; lower left shows the sample values; upper right shows the empirical sampling distribution, with one sample highlighted as a black dot; middle right shows the sample values in the highlighted sample; lower right shows the sample size plot for the highlighted sample. At time shown above, S was viewing animation of the sample size increasing from 25 to 100.

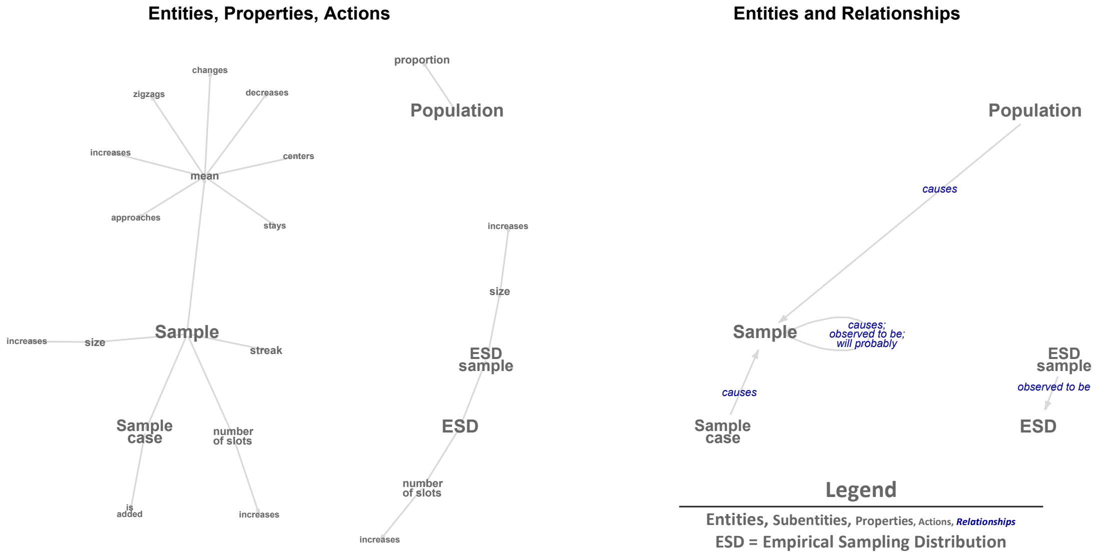


Figure 4.48. Summary of coded entities, properties and actions (left) and entities and relationships (right) for the Growing Many Means: 0-1 activity for participant S.

When asked specifically about what they noticed happening on the distribution plot, they seemed to only note the movements of the highlighted mean as sample size grew, e.g. the comment "it kept shifting to the left, and then it'd sometimes go a little bit to the right" (4.1200–4.1202). When asked a follow-up question about the distribution overall, all S offered was "um, it's a wider distr—like, there's more—there's more options" (4.1209–1210), again noting the increase in the number of possible values for the mean, and again confounding that increase in possible values with terms for the variability of the distribution.

However, S's attention to the moving mean on the top sparked an expression of the connection between swamping and the number of slots:

> I: So what do you notice happening so far?
>
> S: Um, it's moving less and less. [OK] At least on the top graph, like it—because there's more—different, like, fraction numbers that you can get, um, [Hm] so it's definitely—there's l—it's not moving quite as much back and forth.
>
> I: Okay. Um, so tell me a little bit more about that. You're saying it's not moving as much back and forth because there's more fraction numbers than y—that you can get?
>
> S: Yeah, just cuz there's more options—there's more combinations for each thing, so, like, say, like you have only four—like you have—you can choose four blocks, like that's—like you have one out of four chances [mh], like, or like—it's like 25% increments, or 0.25 increments? While, like, if you only have, like—um, like three, it'll be like 0.33. [OK] And so it just kind of keeps on getting smaller—like the spots—like, in between those numbers. [OK]
>
> I: Um, and, um—so that's—you're using that to explain that graph or this graph?
>
> S: The top graph. [OK] Um, just because it's not moving as much, so. [OK] It's kind of like—if it do—like if one block of the—of a different color gets drawn, like it's not gonna it's gonna mo—it's gonna less in one direction, um, than it would, say, like, uh—in if you only had, like, two blocks. Because there's—like a higher, like, percentage of blocks in it, so, it's not gonna affect it quite as much. (4.1322–1365)

Although S had previously expressed elements of the swamping mechanism with sophistication and subtlety, they here were only able to draw a connection between the increasing number of slots and the decreasing movement of the mean as the sample size increases. Because there were less "spots between those numbers", S explicitly connected to swamping by noting that if a different color is drawn, it's "not gonna affect it quite as much". This, however, demonstrated a rather limited connection of swamping to heaping: S used the word "combinations" here, but did not appear to be referring to the different possible permutations in the way that they had used the word before. Instead, and for the remainder of the interviews, this word now was a synonym for the number of slots.

4.5.6 Growing Many Means: 0-1-1

After adding another 1 into the sampler, S's reasoning about the population mean again became vaguer, with only the recognition that the probability of getting 1s would be higher and that sample means would be closer to 1 than before. After some pointed questioning, S did provide a little bit more reasoning about the shape of the ESD, and they were able to observe multiple extreme means converging on the top plot, yielding more *ESD* and *ESD Sample* codes (Figure 4.49, left) than had occurred in the previous activity. Again, however, S focused on causal relationships in individual samples and S's reasoning about ESDs was only observational (Figure 4.49, right). The code *ESD* was applied to S's reasoning about any dotplot of means, including the segment of this activity where S viewed only six extreme means as sample size grew. This was not technically an empirical sampling distribution of a 0-1 variable, but it was still coded as ESD because S's reasoning about multiple simulated means was what was of interest, not the technical accuracy of the distribution label.

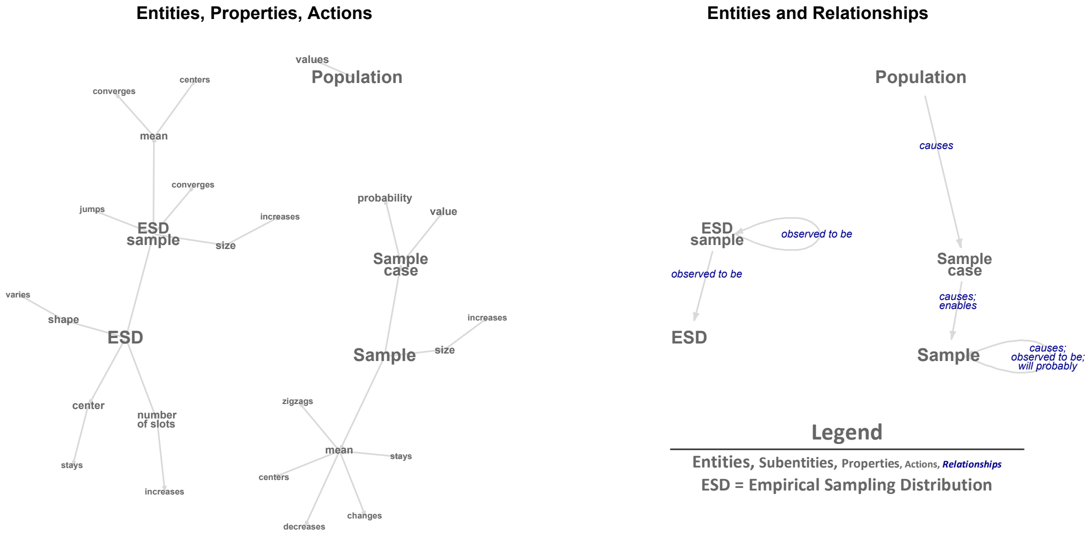
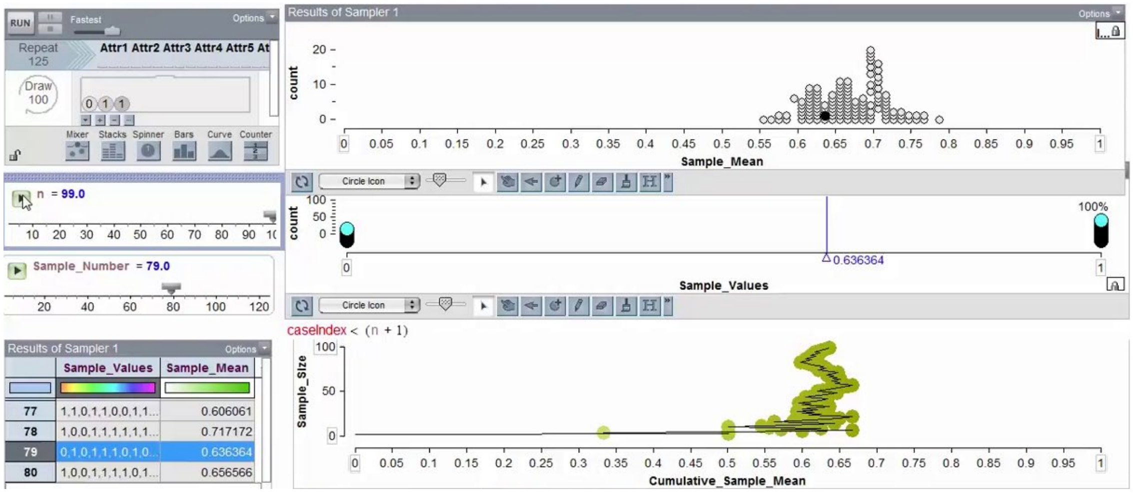


Figure 4.49. Summary of coded entities, properties and actions (left) and entities and relationships (right) for the Growing More Means: 0-1-1 activity for participant S.

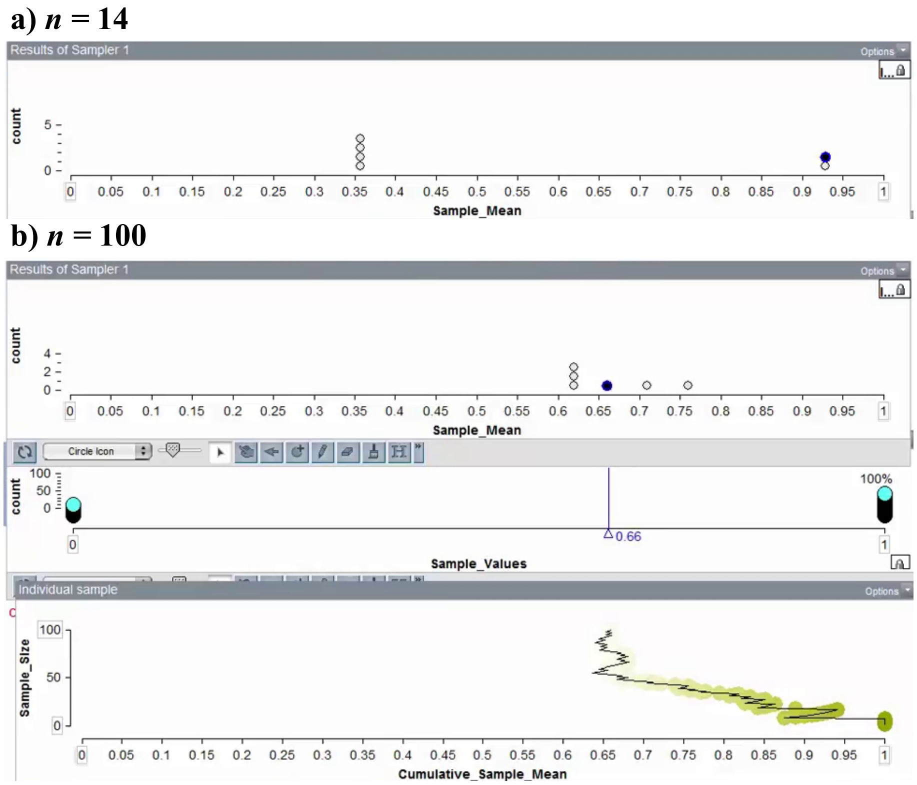
Some more evidence of S's lack of use of the ESD emerged in this section, however, at least when cued to watch one sample in particular. S watched one sample increase from 4 to 99, with one sample mean highlighted on the ESD and with the sample itself and its sample size plot displayed in linked representations (Figure 4.50). S noted that the particular mean was centered between 0.6 and 0.65, and all they noted about the top graph was that it was shaped "like a bell curve […] centered around the same area" (4.1506–1508), even though the ESD was centered a bit higher. S then was shown a plot of just the most extreme means at sample size 14 (Figure 4.51a) and was asked what they expected to happen as the sample size grew, and S responded, "Um, they'll definitely still center at some point around like 0.6 and 0.65" (4.1517–1520). Although it is possible that they were basing the range 0.60–0.65 on a ballpark estimate of the mean of the ESD in Figure 4.50 by noting that the distribution was slightly skewed and thus the mean would be to the left of the mode, S's general preference for modes as a measure of center makes this explanation less plausible. Instead, it is likely that S chose this range because it was the same range that they observed for the individual sample they had viewed previously. Therefore, it seemed that S's judgment of what would happen was based just on the single sample and not on the ESD.



Results of Sampler 1

	Sample_Values	Sample_Mean
77	1,1,0,1,1,0,0,1,1...	0.606061
78	1,0,0,1,1,1,1,1,1...	0.717172
79	0,1,0,1,1,1,0,1,0...	0.636364
80	1,0,0,1,1,1,1,0,1...	0.656566

Figure 4.50. TinkerPlots™ setup and results for Growing Many Means: 0-1-1 for participant S. Upper left shows 1 twice as likely as 0 on a single draw; middle left has sliders for controlling sample size and which sample is highlighted on the top right and shown in the middle and lower right; lower left shows the sample values; upper right shows the empirical sampling distribution, with one sample highlighted as a black dot; middle right shows the sample values in the highlighted sample; lower right shows the sample size plot for the highlighted sample. At time shown above, S was viewing animation of the sample size increasing from 4 to 99.



a) *n* = 14

b) *n* = 100

Figure 4.51. Growth of extreme sample means in TinkerPlots™ for participant S. The plot in a) showed the most extreme 6 sample means at sample size 14, with a sample means at 0.93 highlighted. The plots in b) show what S watched as the sample grew to sample size one hundred, with the same six means on top, the distribution of the sample that corresponded to the highlighted mean in the middle, and the corresponding sample size plot on the bottom.

However, when S viewed the plot of the extreme means at sample size 14 (Figure 4.51a) grow to sample size 100 (Figure 4.51b), S did attend to the overall movement of those six extreme means rather than only looking at the highlighted one:

> S: Um, so for the top graph, all of them started like slowly coming together. They would jump around a little bit, but they all slowly, um, started going towards like the point six, point six five area, or point seven, I guess. [mh] And then for the bottom part, it zigzags but it kept c—shifting more and more towards the left, [OK] um, centered around point six five it looks like.

> I: Okay. And, uh, one last time, so what, uh—why do you think that is? Why did it do that?
>
> S: Um, probably because it had more, uh, white in the mix, um, and there's a higher, like—just because it started off with, um, all blacks that doesn't ma—necessarily mean that, uh—or, like, quite a few blacks, that—it doesn't mean that it can't go to the left, um, and it'll know definitely, like—each white block will, like, affect it too, so. [mh] It won't—it'll be close to one, but not exactly one. (4.1534–1559)

Perhaps watching a handful of means grow was more trackable than viewing the whole distribution, allowing S to see a broader picture than 1 mean without overwhelming them with 100 means simultaneously. S was able to view the means "slowly coming together" and this was a glimpse of a more global view of the process rather than just a single sample. However, their final explanation of why the mean shifted over from 0.93 to 0.66 lacked many of the elements of mechanism, probably because S did not have a sense of the population mean for the 0-1-1 situation. They simply noted that it can still get white blocks, and that "it'll be close to one, but not exactly one."

4.6 Post-Interview

The final interview had the most sections of any interview, encompassing 5 post-questions, a revisitation of the 6 pre-questions, a comparison of the new and old responses on the pre-questions, a short assessment of S's perception of connections between two post-questions and the sample size and permutations plots, and general closing questions about what S observed. S often was reasoning about various aspects of the sample, but drew on a great diversity of other entities. Notably, there was not a presence of the TSD in this interview and S made only a very minimal use of the concepts presented in Growing Possibilities (Table 4.6). Much of S's reasoning was swamping reasoning that attended to the effect of a single case, hence the prevalence of *Sample case.* As noted below, S attended to the number of people or games played in an average day at the casino, which did not fit clearly into any existing entities.

Table 4.6
Presence of Entity, Property, and Action Codes that Occurred in More than One Section of Interview 5 for Participant S

Entity	Property	Action	Presence in Each Interview Section
			Geology Bottles Exam Coin Working Hospital Referend Candy Batting PO Casino Comparison Represent. Closing
ESD			_ _ ■ _ ■ _ ■ _ ■ _ ■ ■ _
ESD	proportion of upper extreme		_ _ ■ _ ■ _ _ _ ■ _ ■ ■ _
ESD	range		_ _ _ _ _ _ ■ _ _ _ ■ _ _
ESD	size		_ _ _ _ ■ _ _ _ ■ _ ■ _ _
ESD sample			_ _ _ _ ■ _ _ _ ■ _ ■ _ _
ESD sample	size		_ _ _ _ ■ _ _ _ ■ _ ■ _ _
Population			_ ■ _ ■ _ ■ _ _ ■ ■ _ _ _
Population	mean		_ ■ _ _ _ _ _ _ ■ ■ _ _ _
Population	proportion		_ _ _ ■ _ ■ _ _ _ _ _ _ _
Sample			■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ ■
Sample		varies	_ ■ _ _ _ _ _ ■ _ _ _ _ _
Sample	mean		■ ■ _ _ _ _ _ _ ■ ■ ■ ■ _
Sample	mean	changes	■ _ _ _ _ _ _ _ ■ ■ _ _ _
Sample	population ratio		_ _ _ ■ _ ■ _ _ _ _ _ _ _
Sample	proportion		_ ■ ■ _ ■ _ ■ ■ _ ■ ■ ■ _
Sample	proportion	changes	_ ■ ■ _ ■ _ ■ ■ _ _ ■ ■ _
Sample	proportion	varies	_ _ ■ _ ■ _ _ _ _ _ ■ ■ _
Sample	sampling		_ _ _ ■ _ ■ _ _ _ _ _ _ _
Sample	size		■ ■ ■ _ ■ _ ■ ■ ■ ■ ■ ■ ■

Sample	size	increases	_	■	_	_	_	_	_	_	_	_	_	■	_
Sample	streak		■	_	_	_	_	_	_	_	_	■	_	_	_
Sample	values		■	■	_	_	_	_	_	_	_	_	_	_	_
Sample case			■	_	■	_	■	_	■	■	■	_	■	■	_
Sample case	value		■	_	_	_	_	_	_	_	■	_	_	_	_
Slot Machine			_	_	_	_	_	_	_	_	_	■	■	_	_
Slot Machine	number of people		_	_	_	_	_	_	_	_	_	■	■	_	_

Note. A symbol of _ indicated that the code did not occur within that interview section, whereas ■ indicated that the interview section contained the code. PO = Post Office, Referend = Referendum, Represent. = Comparing Representations, Comparison = Comparing Old and New Pre-Questions.

4.6.1 Post-questions

S did not seem to particularly latch onto a strong understanding of the heaping mechanism during prior activities, and thus it was not surprising that their reasoning on the post-questions did not incorporate the TSD and was dominated by examination of different features of what affected an individual sample mean, including the swamping-type explanations that S used throughout the study.

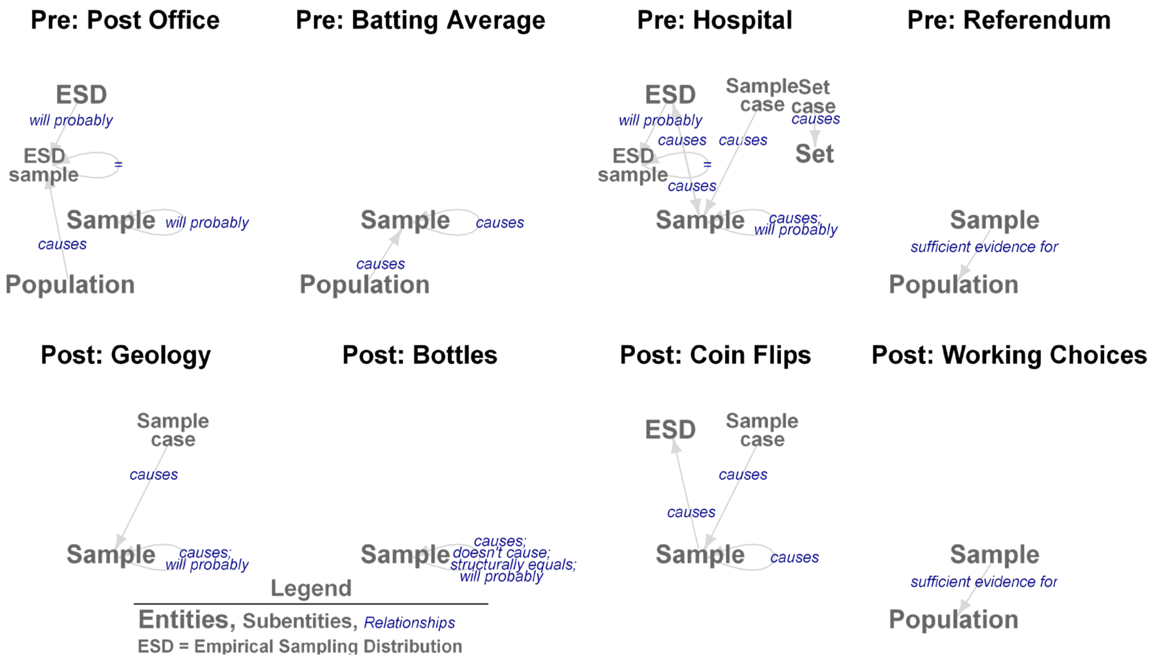
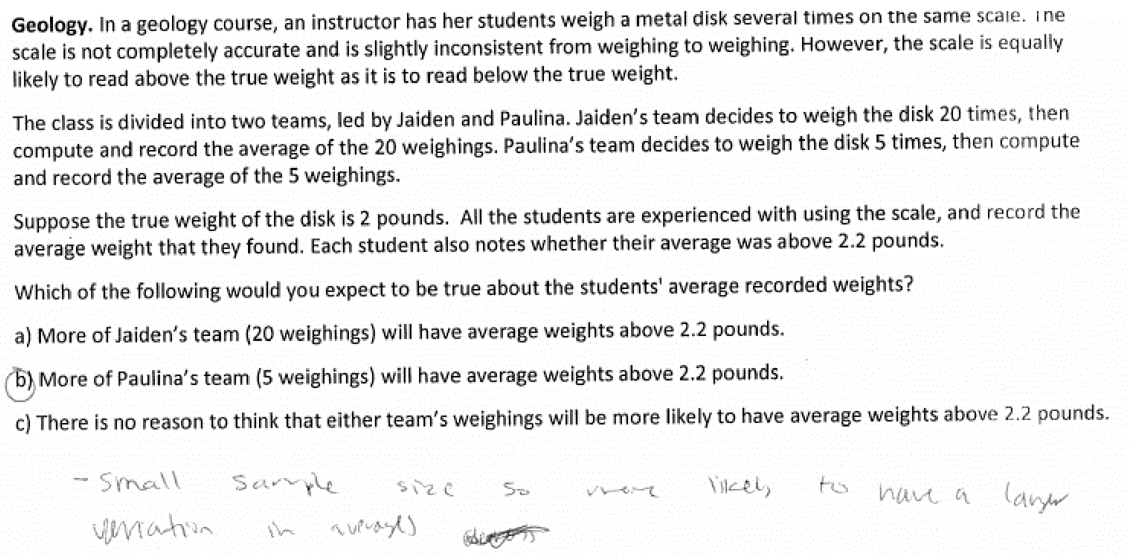


Figure 4.52. Entity and relationship summary for post-questions (bottom) and the corresponding isomorphic pre-questions (top) for participant S. S did not identify any entities or relationships for the Exam Preparation post-question.

S's reasoning about entities and their relationships was fairly similar on the post-interview as it was to isomorphic questions on the pre-interview (Figure 4.52). Even though the Geology post-question was phrased in terms of ESDs, S did not focus on this aspect, perhaps because the *ESD size* was not mentioned, whereas *ESD size* played a large role in the Post Office pre-question. Coin Flips showed a similar lack of focus on the *ESD size* as compared to the Hospital pre-question, and again no *ESD size* number was given. Geology was a repeated measurements problem whereas Post Office was sampling from a fixed population, and S may have reasoned differently about these contexts (cf. Konold & Harradine, 2014). S attended less to the population process in Bottles than in Batting

Averages and S's focus was more on the order of responses, perhaps because of activities involving permutations, as discussed below.



Geology. In a geology course, an instructor has her students weigh a metal disk several times on the same scale. The scale is not completely accurate and is slightly inconsistent from weighing to weighing. However, the scale is equally likely to read above the true weight as it is to read below the true weight.

The class is divided into two teams, led by Jaiden and Paulina. Jaiden's team decides to weigh the disk 20 times, then compute and record the average of the 20 weighings. Paulina's team decides to weigh the disk 5 times, then compute and record the average of the 5 weighings.

Suppose the true weight of the disk is 2 pounds. All the students are experienced with using the scale, and record the average weight that they found. Each student also notes whether their average was above 2.2 pounds.

Which of the following would you expect to be true about the students' average recorded weights?

a) More of Jaiden's team (20 weighings) will have average weights above 2.2 pounds.

[✓ b)] More of Paulina's team (5 weighings) will have average weights above 2.2 pounds.

c) There is no reason to think that either team's weighings will be more likely to have average weights above 2.2 pounds.

- Small sample size so more likely to have a larger variation in averages

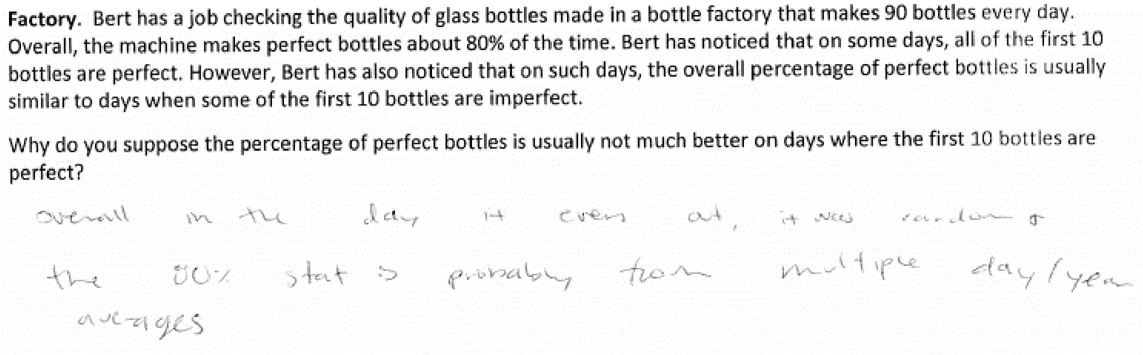
Figure 4.53. S's response to the first post-question, the Geology problem (Brown, 2016; Well et al., 1990).

Despite the similarities between the pre- and post-questions, there were intriguing hints of new reasoning in the first two post-questions. In reasoning about the Geology problem (Figure 4.53), S at first gave some swamping reasoning about the "effect" that an individual weighing would have on the average, but supplemented it with reasoning that had hints of both balancing and heaping:

> Like, if they have one at, like—for some reason like—gave, like, a two point five, and then they also, like—and their—keep on like staying more above two than under two, [OK] um, for those like next five times, then if their average is definitely going to be higher, but if they have like 20 times, like, they might have five in a row that are above 2.2, but they

> might also have five below 2.2. [OK] So, five doesn't really give like a good, um, mixture. Like, it d—it just isn't a big enough sample size, so. (5.87–99)

As they have done several times throughout the interviews, S reasoned based on the possibility of streaks of values, which could have provided a way of chunking elements of the mechanism for easier processing. Here, S discussed how a small sample may have a streak all above the mean, but in a larger sample, they could have also had a streak below the mean that would balance out to the true mean. S then said that the sample of five did not give a good "mixture", a term that they first used when describing the permutations of building blocks, at how there were many possible ways of mixing the building blocks at the mean of 0.5 of a 0-1 variable. This suggests S may have made an unanticipated connection between heaping and balancing: As sample size grows, possible means heap near the true value because there are more opportunities for mixing, or balancing out, to be a mean that is near the true mean.



Factory. Bert has a job checking the quality of glass bottles made in a bottle factory that makes 90 bottles every day. Overall, the machine makes perfect bottles about 80% of the time. Bert has noticed that on some days, all of the first 10 bottles are perfect. However, Bert has also noticed that on such days, the overall percentage of perfect bottles is usually similar to days when some of the first 10 bottles are imperfect.

Why do you suppose the percentage of perfect bottles is usually not much better on days where the first 10 bottles are perfect?

overall in the day it evens out, it was random &
the 80% stat is probably from multiple day/year
averages

Figure 4.54. S's response to the second post-question, Factory (Nisbett, Fong, Lehman, & Cheng, 1987).

Perhaps using similar line of reasoning, S explicitly invoked the building block activity to explain regression to the mean in the second post-test question (Figure 4.54):

> S: Um, and so you really have to look at like the overall day rather than the first ten, cuz, like, it's kind of like with the blocks how you could, like, have two white blocks in the beginning, and you have two black box—black bo—blocks in the beginning, but it's it—still the same percentage of getting like two black blocks in the beginning, and then two white blocks at the end, so.
>
> I: Okay. Um, so I didn't quite follow what you just said about the—the blocks. Could you tell me a little bit =more about that?=
>
> S: =Yeah, so like= when you draw, say, four blocks, you could have two blocks—you could draw two white first, and then you could draw two black. [OK] So then, it's a 50/50, like you have—like out of the four blocks that you got, you got two white, two black. But then you could also, like, if you did another round, you could have got like, say, two blacks, and then two whites, [Hm] and it's still like overall the same percentage, it just depends on like the—um, like the—like the, um, way that you grabbed them. Like it's—overall, like, at the end—it didn't—it doesn't really matter, you still have like those—that, like, the same amount of, like, white and black. [OK] So it just kind of—but like the order of them changed, so. (5.162–195)

Something that S appeared to have gotten out of the Growing Possibilities: 0-1 Building Blocks activity was that different orders of drawing the blocks could lead to the same mean. So, just like one could either draw 0-0-1-1 or 1-1-0-0 and still end up at the same mean of 0.5, having the first 10 bottles all be perfect could be part of a total sample of 100 that still is at 80%. Apparently, S was arguing that observing 10 perfect bottles would be irrelevant to the mechanism because the building block situation showed that the order did not matter. Especially in the case of the building blocks, this reasoning is not quite normative. If the first two blocks were observed to be 1s, the most probable sample mean of four blocks would be 0.75, whereas if the first two blocks were observed to be 0s, the most probable sample mean would be 0.25. Therefore, the first blocks drawn do affect the likelihood of the eventual mean. However, this attention to order—certainly an important part of the mechanism—could prepare S for further exploration and development of understanding of the heaping mechanism. This episode demonstrated both the promise and peril of attempting to build deeper conceptual understanding, since the additional exposed elements of the mechanism can both support long-term learning but also lead to new and unexpected misconceptions.

Despite the richness of S's engagement with mechanism throughout the interviews, S had apparently a relatively impoverished mechanistic understanding of the production of a rerandomized *p*-value, which by that point had been covered in the CATALST course (Zieffler & Catalysts for Change, 2017). S had been willing to engage with many sample size situations and had always quickly developed thoughts and explanations, but not with the Exam Preparation problem, which asked them to choose which randomization distribution they would expect for a comparing two groups problem with a larger sample size. S described how they were not able to follow the procedure that they knew in class:

> So like usually it's like—okay we have like the graph, and then we know, um, we have like that benchmark. So it'd be, like, the mean difference of like five, and then we would have that like count [Hm] of like how many are more extreme or like equal to that. [mh] And then we would do that over the, um, like number, the sample size. [...] I feel like—it just—like I just don't under—I feel like I just like personally like there's not enough information given [OK] for me to even like give like an educated guess, because [OK] like you don't have any—like you're basically told like—okay, so like for this one, it's like okay so you have a sample size of 200, you did the 500 randomized trials, and your mean difference was 5 but you don't have any other information besides that [/mh?/]. Besides—the only other information you're given is there was a different sample size, and this was—these were the results but you don't know, like, what really like—what the ex—results were for them, so I feel like there's no way to like accurately say like yeah this is probably what the p-value could be, cuz, like that—that's just like what I'm thinking. (5.378–434)

S knew how to calculate the *p*-value given the randomization distribution, but did not have any way to reason about the situation without “what really the results were”, or a way of reasoning about what they would expect the randomization distribution to look like at a different sample size. This was, admittedly, a much more complicated data-generating mechanism than any of the other problems given in the two interviews, and S felt like they could not “even give like an educated guess”, because the added complication of re-randomizing means made the entire mechanism opaque to them.

4.6.2 Pre-questions, Revisited

S was given the opportunity to re-examine the pre-questions twice, first on new pieces of paper, and second to compare their initial responses on the first interview with the present interview. S’s reasoning about the pre-questions were similar on the post-interview, and they attended to many of the same entities and relationships between entities (Figure 4.55). The biggest differences in entity and relationships was in the Batting Average problem, where S attended more to the causal relationship between *Sample case* and *Sample* as compared to their somewhat vaguer reasoning on the pre-interview, and they had an added sense that a smaller sample size *enables* the mean to vary and a large sample size *prevents* it from doing so. The Post Office problem also included more swamping reasoning about the impact of a sample case on the sample mean. In the Casino problem, where S assigned a stronger causal role in the post-interview to the original context of the number of people they thought would play the slot machine. The entities and relationships in the Hospital, Referendum, and Candy problems were quite similar between the pre- and post-interviews.

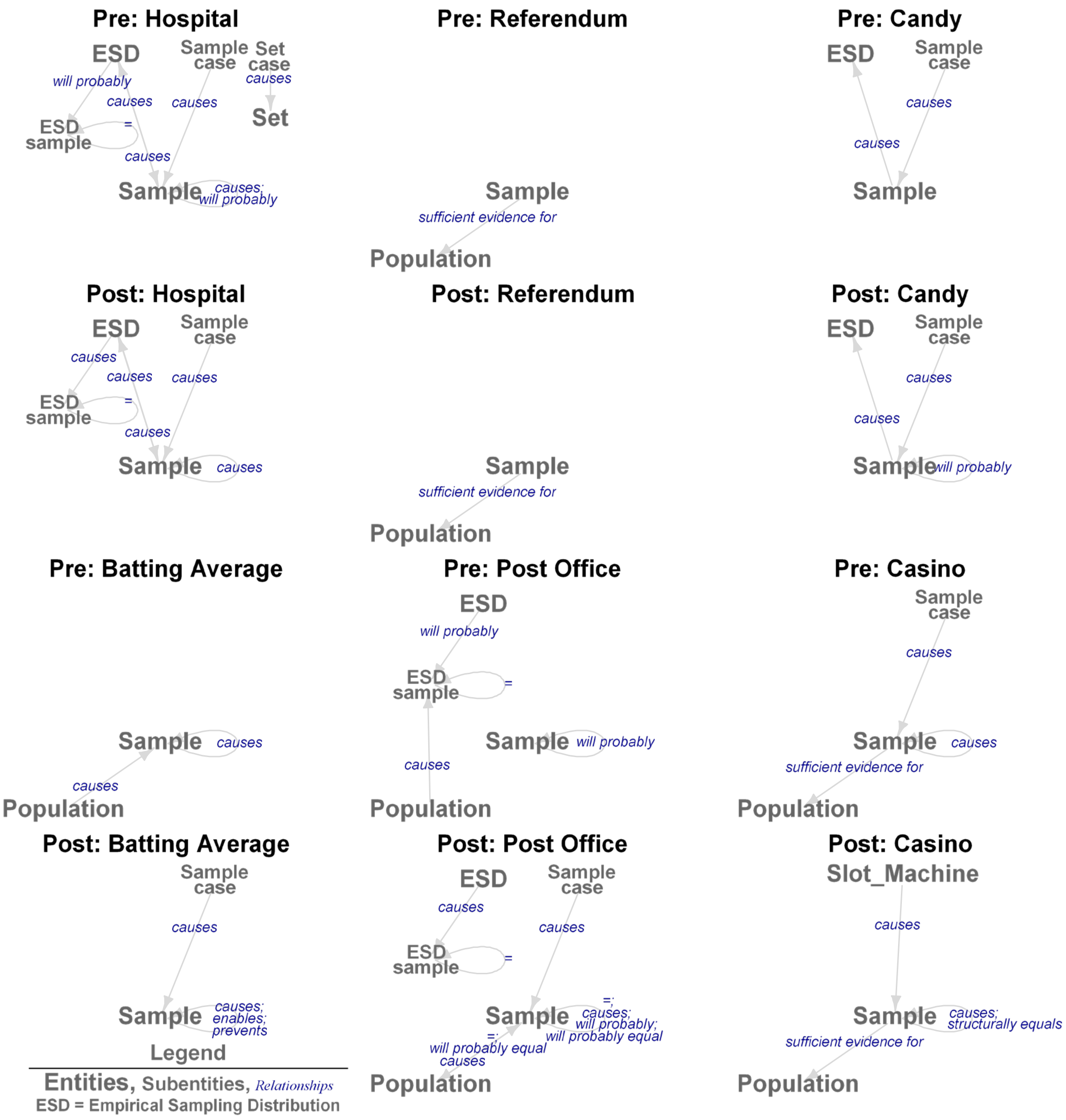
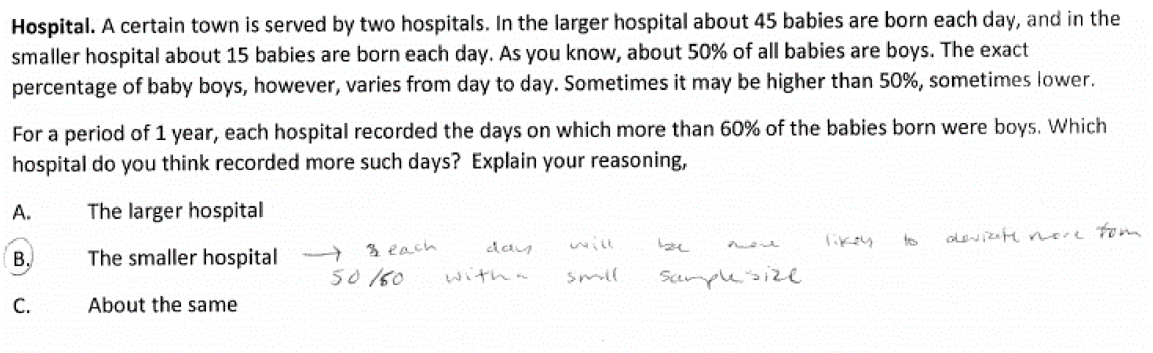


Figure 4.55. Entity and relationship summary for pre-questions, both in the pre-interview (rows 1 and 3) and in the post-interview (rows 2 and 4) for participant S.



Hospital. A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50% of all babies are boys. The exact percentage of baby boys, however, varies from day to day. Sometimes it may be higher than 50%, sometimes lower.

For a period of 1 year, each hospital recorded the days on which more than 60% of the babies born were boys. Which hospital do you think recorded more such days? Explain your reasoning,

A. The larger hospital

[✓ B.] The smaller hospital → % each day will be more likely to deviate more from 50/60 with a small sample size

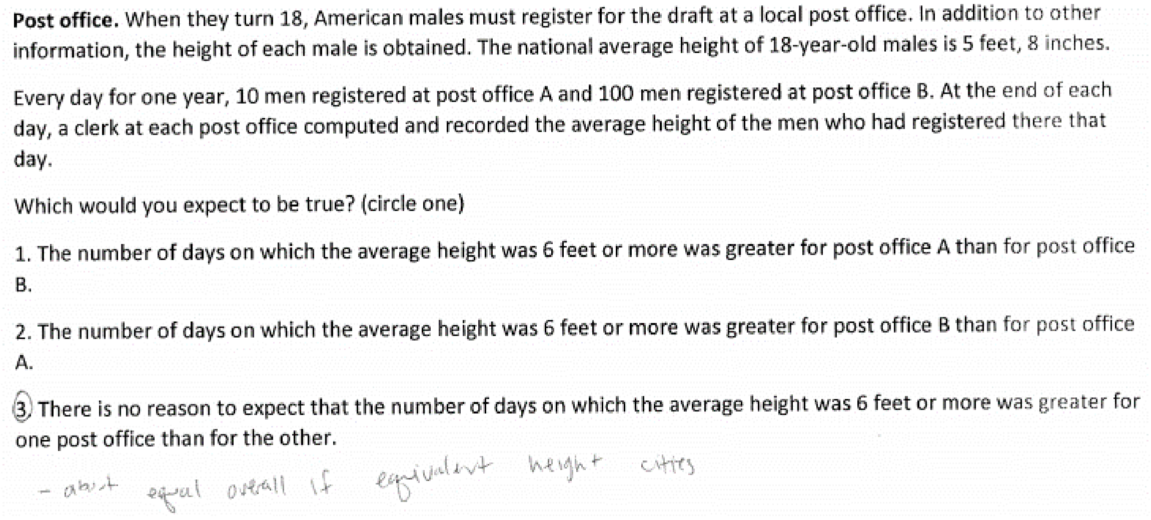
C. About the same

Figure 4.56. S's response to the Hospital problem (Kahneman & Tversky, 1972) in the post-interview.

The most dramatic change, however, from S's pre- to post- responses to these same questions was in the Hospital problem, when S correctly chose the smaller hospital (Figure 4.56) in contrast to their choice of "about the same" in the pre-interview (Figure 4.2). After initially stating that they would choose the smaller hospital, because one baby boy would change the proportion "quite drastically" (5.671), S briefly changed their answer to "about the same" in line with their pre-interview response, but then course-corrected back to choosing the smaller hospital:

> Actually, I'm going to change my answer. Uh, it's—because I forgot that was over a period of one year. [OK] Um, I would probably say that they're about the same. [OK] Yeah. I would say p—m—{which hospital do you think recorded more such days...} Uh. No. I'm going to stay with my—smaller hospital. [OK] Yeah. Cuz it's like—I j—I going to change it, but then when I read, like, just which hospital do you think recorded more such days, so overall I think they would probably gonna be around 50 each, just because 365 days [Hm]—that's a lot of days that you can get data for. [mh] But looking at it individually, and which days, like—you know they could have 10—like they could have like seven baby girls or something. [mh] Or like eight—or like ten baby girls. Um, and that's definitely gonna, like, increase or decrease the percentage. [mh] And—but overall, like, they could also have like the same for boys. So I would say still the smaller hospital, just because it's more likely to deviate, [OK] um, while like the—the bigger hospital probably has more of like—it doesn't go up and down quite as drastically. (5.681–706)

Here S's thinking was on the border between thinking about the problem as asking about which hospital would be close to 50 "overall"—presumably if the 365 days were pooled together into one big sample—and thinking about each day "individually", looking at what would be more or less likely within a single day. These were two conflicting strategies for recasting this problem that asks about ESDs to be about single samples, a tendency that Well et al. (1990) also observed. The "big sample" strategy leads to an incorrect answer, whereas the "individual day" strategy that S came to at the end allowed them to apply their mechanistic reasoning about how each case will affect the proportion more strongly in the small sample. When comparing their post-response with their pre-response, S commented that when they originally viewed the problem they were considering "whatever 45 times, um, 365 is, like, um, looking at like that number, [OK] compared to like the individual days" (5.1238–1240), which is consistent with the interpretation that they were using the big sample strategy. It is not clear why S never used the big sample strategy on the Geology or Coin Flips problems, but some of the issue may be that the Hospital and Post Office problems are both phrased in terms of "over the course of a year", which may have encouraged S to think of them as growing a big sample rather than gradually accumulating an ESD. Furthermore, as noted above, neither the Geology or Coin Flips problems explictly gave the *ESD size*. The Batting Average problem, in contrast, is also over the course of a year but is appropriately considered as a growing sample.



Post office. When they turn 18, American males must register for the draft at a local post office. In addition to other information, the height of each male is obtained. The national average height of 18-year-old males is 5 feet, 8 inches.

Every day for one year, 10 men registered at post office A and 100 men registered at post office B. At the end of each day, a clerk at each post office computed and recorded the average height of the men who had registered there that day.

Which would you expect to be true? (circle one)

1. The number of days on which the average height was 6 feet or more was greater for post office A than for post office B.

2. The number of days on which the average height was 6 feet or more was greater for post office B than for post office A.

[✓ 3.] There is no reason to expect that the number of days on which the average height was 6 feet or more was greater for one post office than for the other.

- about equal overall if equivalent height cities

Figure 4.57. S’s response to the Post Office problem (Well et al., 1990) for the post-interview.

After S corrected their reasoning about the Hospital problem, and providing relatively similar reasoning as the pre-test on the Referendum, Candies, and Batting Averages problems, S again chose the equal option for the Post Office problem (Figure 4.57) as they had on the pre-interview. It was not clear why they reasoned about this differently than the Hospital problem, although perhaps responding to the Batting Averages problem primed them to think from a big sample perspective, since order effects on sample size problems have been observed in prior literature (Brown, 2016; Sedlmeier, 1998). Their reasoning was more ambiguously phrased, but did have some suggestions of a big sample perspective:

> So with the number of people that are six feet or over, um, the average for—of height for each post office is gonna be probably the same [OK], um, just overall. It'll be very close to five eight, but I don't think there's any reason to think, um, that there's gonna be a higher amount of days that, like, the average height was over six feet for one or the other. I think,

> like, ten men might—like the post office A might have more numbers, but I feel like over a year, they're probably gonna be about the same overall for, um, which is gonna be over six feet, so. (5.1069–1083)

As they did in the pre-interview, they clearly stated that they would expect the small post office "will deviate more" (5.1014), but they expected them to be "about the same overall". The statement "it'll be very close to five eight" suggest that "it" may be the mean of the large sample, even though they use the language of the problem and say that the number of days will be similar. When later comparing their old and new responses, S recognized some dissonance with their changed response on the Hospital problem:

> Um, I think this one—um, I almost, like, kind of corrected myself, I think from the Hospital for this one. Um, I didn't really think about it, probably. But, uh, because I talked about it—I specifically said the amount of days collected is probably going to be the same, um, so I think that I went—uh, went by it the same way *[as the pre-interview]*. Like, the, like, overall like in a—like a week, the percentages might be drastically different, but if you look at the specific days I feel like it'd probably be pretty equivalent. So, I think I went about it the same way. (5.1315–1328)

S recognized that there was something in common with the Hospital problem, but for some reason here stated that in a *week* that the percentages may differ but at "the specific days" would be the same, perhaps because of their emphasis that the overall average would be similar. Although their reasoning during the first presentation of the problem in the post-interview seemed to be driven by big sample reasoning, in this comparison it may more be that they are trying to justify the answer that they had already given, since it seems like they were reverse-engineering their response since they are attending to what they "specifically said" on the post-interview, and they had never mentioned anything about a "week" before or why that would differ from "specific days".

4.6.3 Comparing Representations

So far, S had not been asked to specifically link the representations introduced during Growing Certain and the textual problems in the pre- and post-interviews. When prompted with sample size plots (from Cat Factory 2) and permutation plots (from Growing More Means: 0-1-1) deliberately chosen to be non-isomorphic to the chosen textual problems (Geology and Coin Flips), S was unsurprisingly able to work more with the sample size plot. Even for the permutation plots, S mainly focused on how different orders would lead to the same sample mean, similar to their reasoning about the Bottle Factory problem, which led to a prevalence of codes regarding the sample rather than any focus on the ESD (Figure 4.58, left). S provided a sophisticated description of the changing impact of a sample case on the sample mean as sample size increased for the sample size plot, but for the permutations plot simply noted the fact that the *Sample order* could differ despite having the same sample mean (Figure 4.58, right).

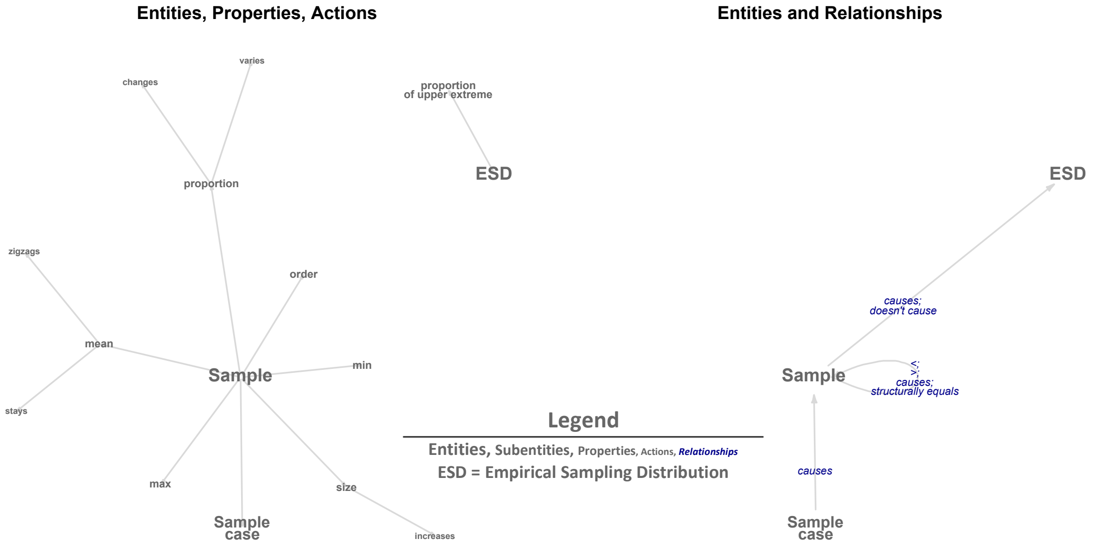
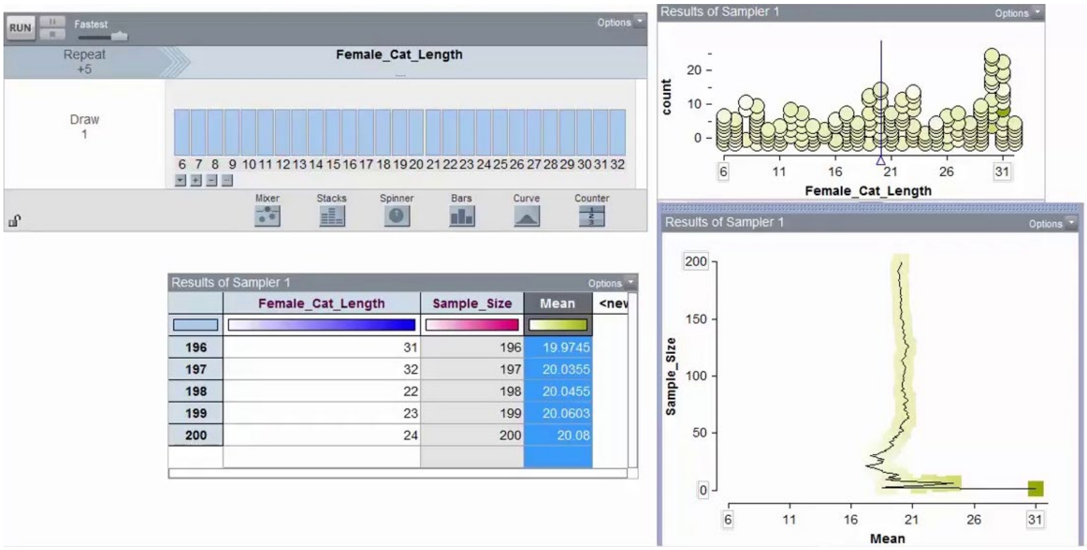


Figure 4.58. Summary of coded entities, properties and actions (left) and entities and relationships (right) for the Mystery Machine # activity for participant S.



Results of Sampler 1

	Female_Cat_Length	Sample_Size	Mean	<nev
196	31	196	19.9745	
197	32	197	20.0355	
198	22	198	20.0455	
199	23	199	20.0603	
200	24	200	20.08	

Figure 4.59. S’s Cat Factory 2 display that S examined during Comparing Representations to find connections with the Geology and Coin Flips problems.

S displayed lots of evidence of mechanistic reasoning when comparing the sample size plot in Cat Factory 2 with the coin flips problem, perhaps because it reflected a familiar process:

> It's gonna like what—the more trials you do it, the more it's gonna really stay at that fifty-fifty overall. Um, while like in the beginning if you only do they say like five flips, it's gonna go up and down, back and forth, quite a bit. Um, it's gonna be quite zigzaggy at the bottom, just like how this one is, but then as soon as you get around like that 50 it kind of really stays, um, basically almost, like, I would say vertical, and it doesn't really zigzag around. (5.1473–1485)

Although S has shown similar reasoning in the Growing Certain activities, this is among the most confident and clear statements of the decrease in zigzagging, with statements of “gonna” and “really stays” and few of the qualifying words S had frequently stated before (e.g. “maybe” or “probably”). S seemed confident about what they would expect with a large number of coin flips and saw this representation as supporting that. In contrast, their reasoning about the Geology problem when viewing this same representation was much more vague:

> Um, I mean it's kind of like—if you look at—if the—the scale... if you think about the scale like kind of in the way of the cat length, um, it's gonna go about even on the same side, so like even that 21 isn't between like the 6 and the 32, um, at one point, like, the average is gonna be somewhere in between whatever two, like ways that it goes. Like, plus or minus, say like, 5 on like the—the scale. Um, so it ma—like, when you look at the graph, um, you know, whatever the—the range of that—that scale is probably gonna probably be, the average is gonna be somewhere in the middle, so no matter—even though like they all have a likelihood of like being the same, um, it's gonna like that average is gonna be in the middle. (5.1398–1415)

S did not talk at all about what happened as sample size increased. Instead, they appeared to be predicting that the measurement error for the Geology problem would be be symmetric, just as the Cat Factory 2 production process was symmetric, and that the mean would therefore be “in the middle” of the range. Perhaps this was salient to them because they originally had not expected their Cat Factory 2 process to settle down at all and observed it settling down to the midrange in the second interview.

4.6.4 Interview Closing

The interview closed with several questions asking what they noticed, what they thought the problems were getting at, what they noticed changing, and how they would explain the relationship between sample size and sampling variability. Although other participants provided extensive mechanistic reasoning during this segment, S did not even mention sample size until pointedly asked a question regarding the relationship between sample size and sampling variability. S commented on general features of problem solving, including the importance of not “overthinking”, and reflected that “my reasoning stayed pretty much the same” throughout the study (5.1598).

The only mechanistic code during this closing section, *a large sample size causes rare values to be likely*, occurred when S was asked how they would explain the relationship between sample size and sampling variability which displayed none of the mechanistic reasoning that S had displayed during Growing Certain:

> Cuz it's like—like thinking about the fruit snacks, like, I will always remember like in track like we would always use, like—we would always get fruit snacks. And if I got the Scooby Doo ones like everybody, like for some reason a lot of our favorite like fruit snack it was, um, like the blue Scooby Snack, [Hm] and, like, we'd always be like oh I hope I get a lot in these and sometimes [Hm] you wouldn't get any, and you'd be so disappointed. [Hm] Um, but if then, it's like if you look in the big picture, like if you have like a bigger bag, you're gonna be more likely to have that [Hm], or even like if like you have ice cream and you put something in there, and you mix it around. Like you're more likely to have a scoop with like—if you have like sprinkles or something or like brownie pieces, like the more you put in, the more you're gonna like each have, like in every single scoop rather than

> [Hm] like some scoops like may not have any sprinkles or like any pieces of brownie, just because like you don't have as many in there, like [mh], so that would be, like, your sample size. (5.1656–1682)

S had raised this memory of wanting blue Scooby Doo snacks both times when they discussed the Candies problem, but in that context noted that larger bags would be more likely to be near “50/50” which at least was connecting with a population process. Here, it seemed that S was simply noting that a larger bag would be more likely to have any blue at all. The ice cream analogy appears to be describing how if there are more sprinkles in ice cream, a scoop is more likely to pick up those sprinkles. This could map onto sampling variability: if sprinkles are the amount in a substance, a higher population proportion means that a given sample (the scoop) is more likely to have a higher proportion of sprinkles. However, S is not comparing smaller or larger scoops, and so it is hard to follow what it could mean that the sprinkles “would be, like, your sample size”. It may be that S is generally reasoning based on a “more is better” heuristic (Wagner, 2006), and it is striking that this is the only sample size reasoning S displays outside of the contexts and representations given in the interview, given the richness of mechanistic reasoning S displayed in Growing Certain and even in the first pre-question, Hospital, at the beginning of the first interview.

Chapter 5

Discussion

This chapter discusses how the results of the study relate to prior work and addresses the research questions:

1. *How and when are students reasoning mechanistically before, during, and after Growing Certain? How does each element of Growing Certain (prompts, representations, etc.) interact with students' mechanistic reasoning about the Empirical Law of Large Numbers?*
2. *How can we describe conceptual change when it happens in the context of Growing Certain? How do the changes observed relate to the task design, including the representations used, prompts, and social interactions?*

These questions are addressed in the next section, "Characterizing Mechanistic Reasoning". The remaining sections summarize implications for supporting mechanistic reasoning of sampling variability, discuss issues with the mechanistic coding system, draw a few broader teaching implications for statistics educators, and discuss the limitations of the present work.

5.1 Characterizing Mechanistic Reasoning

The first research question concerning mechanistic reasoning is addressed below by examining how S's reasoning related to the categories of mechanistic reasoning described by Russ et al. (2008), and which were the initial proposed codes for analyzing the data (see Chapter 3): describing the target phenomenon, identifying setup conditions, identifying entities, identifying properties and activities, identifying organization of entities, and chaining. The original coding plan also included “analogies”; S’s reasoning using analogies is presented below as a part of discussing the relevant corresponding parts of the mechanism, as appropriate. The second question is also addressed below by examining the changes that occurred in each type of mechanistic reasoning.

5.1.1 Describing the Target Phenomenon

Russ et al. (2008) characterized “describing the target phenomenon” as the most basic level of mechanistic reasoning, and this was the major focus of study for most of the sample size neglect literature reviewed in Chapter 2: How do people reason about different types of sample size problems? S’s descriptions of phenomena associated with increasing sample sizes was already heterogeneous at the pre-interview, and remained sensitive to representations and the population setup throughout all interviews. S’s description of what happened to means or proportions as sample size increases was a delicate balance of attention to the variability at low sample sizes, and attention to the long-run convergence at higher sample sizes. When responding to the hospital problem and discussing a small sample, S noted “it’s gonna be a little bit more varied” (1.277), highlighting variability, and then noting that at a larger sample size that it’s going to be “around that 50 percent” (1.281–282), highlighting the long-run center.

In the second interview, when S explored the situation of two blocks having a 50% chance of being drawn, S's descriptions became increasingly rich as they gained experience with the situation and broadened their representational repertoire: recording the changes in percentages in a physical simulation, drawing a larger number of blocks in a TinkerPlots™ simulation, and finally viewing the sample size plot. At first, they simply noted that they would expect 10 blocks to end up around 50%. After recording the 10 blocks, S was able to notice that the changes were decreasing, which became a point of departure for swamping reasoning. When watching a sample proportion grow in TinkerPlots™, S again noticed the decreasing percentage and then how "it kind of started going more towards […] 50/50" (2.450–453), again shuffling between attending to the variability and then to the center. S's description of the initial variability developed further once they viewed the mean of a 0-1 variable growing in TinkerPlots™. Instead of just noticing the decreasing changes, the sample size plot seemed to support them to see the mean "zig-zagging" at lower sample sizes, and that zig-zagging decreasing as sample size grows. Furthermore, the sample size plot allowed S to notice that often there was a *critical sample size:* a sample size after which there was less movement of the mean and after which things were relatively stable, which S also drew on when describing the sample size plots in the Cat Factory activities.

These developments in the ability to describe the phenomenon of sample size increasing for S appeared to be clearly related to the growing a sample paradigm, and in particular to the affordances of the sample size plot. Although S came into the study with an apparent propensity for swamping, the sample size plot seemed to make aspects of the representation clearly visible. Since this plot showed both the individual contribution of each additional case in the plot, the patterns of “zigzagging” and the eventually tapering into a straight line seemed to be a useful proxy for viewing sampling variability at different sample sizes, and seemed to allow S to describe the phenomenon of sample size growth more precisely. These more precise observations of growing samples allowed S to attend to entities, properties, and activities that they had not attended to previously, leading to deeper chaining about the mechanism.

The relative success of the sample size plot in promoting richer mechanistic reasoning is in contrast to Sedlmeier’s (1999) claim that the way to promote normative sample size reasoning is through allowing people to clearly translate their intuitions about sample size by having them generate sampling distributions through a “flexible urn” model. Although S indeed was using a sampling system similar to that in the simulation in Sedlmeier, it was the specific representation of the sample size plot that seemed to support their thinking more than the visible sampling system, and it did not seem that translating between frequency and sampling distributions was central to their ability to reason about sample size problems. Furthermore, the sample size plot seemed to encourage a deeper understanding of features of the problem than Sedlmeier’s proposed size-confidence intuition. Although S at times had some sense that they expected the mean to converge to the population mean, this sense was not always present, and S’s reasoning usually relied more on the dynamics of swamping and balancing than on increasing similarity of a sample to the population it was drawn from as sample size increased.

However, there is evidence from S's reasoning to support Sedlmeier and Gigerenzer's (1997) contention that people may be much more facile at reasoning about single samples than about multiple samples. When S reasoned most successfully about sampling distributions, it was largely when they were using the logic of what they would expect to happen within a single sample; when examining multiple samples, S usually got confused and conflated the sample size with the number of samples. Especially direct evidence of this was found when S based their reasoning about whether 100 was a large enough sample size for telling whether the casino's claim of 90 cents was accurate in Mystery Machine #2 entirely based on the sample size plot of one sample, and not based on the ESD with sample size 100. In the Growing Many Means activities, S generally ignored *both* the frequency and sampling distributions and focused on the sample size plot and the movement of *one single mean* in the sampling distribution plot. Later, after simultaneously observing one sample and the entire ESD growing to sample size 100, S based their prediction of a different sample again just on the first sample and not on the ESD. Note, however, that this still contrasts slightly with Sedlmeier and Gigerenzer, who proposed that the "frequency distribution" is what was intuitive to people, whereas S seemed to get the most mileage, even in their pre-question reasoning, from reasoning about *changes in the mean or proportion* as the sample size grew, and especially when this was visibly supported by the sample size plot.

S’s relatively impoverished reasoning at the end of the final interview, where they boiled down the Empirical Law of Large Numbers to the idea that “more is better” by analogy that the more sprinkles you put in ice cream, the more likely you are to have them in a single scoop, also raises the question of how much of instruction should be based on relatively basic intuitions such as “more is better” or the size-confidence intuition. When is an intuition productive, and when can it be built on and connected in a way that is useful to people? S’s use of these intuitions did not appear to give much traction, and it might be argued that intuitions are useful if they can be incorporated into a richer, more well-connected and structured mechanistic understanding similar to diSessa’s (1993) characterization of the gradual development of understanding of physical mechanism out of the fragmentary *p*-prims of knowledge.

5.1.2 Identifying Setup Conditions

Although Russ et al. (2008) seemed to cast “identifying setup conditions” at a relatively low place in their hierarchy of mechanistic reasoning, successful reasoning about populations and processes was quite challenging for S throughout. Understanding the “setup conditions” can be a central struggle in understanding probability and statistics, and so this hierarchy may not apply in the same way to this setting. In particular, S appeared to be able to reason with relative success about 0-1 equiprobable situations, and symmetric bell-curved situations, but with less success in the 0-1-1 situation, a uniform distribution, or when reasoning about the relationship between sample size and the distribution of rerandomized trials for a randomization test.

The difficulty that S had with the 0-1-1 situation was somewhat surprising; S struggled to even set up TinkerPlots™, questioning whether the proportions should be 25% and 75%. This lack of clarity about the population process may be partially due to lower arithmetic fluency, which there was other evidence during the interviews. After repeatedly returning to the 0-1-1 situation from a number of angles, S never clearly expressed a sense that there was a population mean of 0.66 or that the sample size plot and ESD would be expected to converge to this number. There was always a sense of vagueness about what that number would be, even though S seemed to think that there would be some convergence, and at times reasoned with sophistication about how the individual sample grew.

When reasoning about continuous distributions, S appeared to attend more to the population mode than to the population mean. Of course, this led to normative conclusions in a bell-shaped distribution (where the mean and mode are the same), but not for a uniform distribution, where S concluded that the mean would move less on each individual draw because of swamping, but that it would not necessarily center in any given place because all the population values were equally likely. Viewing what actually happened with the uniform distribution allowed S to see limits of their existing reasoning and to deepen their understanding of mechanism by bringing in balancing to accompany the swamping reasoning, since swamping alone did not predict convergence. However, this reasoning still did not involve the population mean directly.

The challenges and opportunities brought by these different populations underscore the importance of working with different types of populations and giving people opportunities to reason about growing a sample with different population shapes. Not much variety in population shapes is in most of the sample size neglect literature, excepting that by delMas and colleagues (Chance et al., 2004; delMas et al., 1999, 2006). The latter research was on students' ability to recognize ESDs at different sample sizes rather than growing a sample, however. S had little success reasoning directly about ESDs or in developing understanding of heaping, and this may be both a cause and an effect of S's difficulty connecting population setup conditions to the properties of samples. If S had developed better understanding of heaping, that could have addressed these issues, but arguably heaping is not able to be understood without seeing a strong connection between the population mean and the means of individual samples. To take fuller advantage of heaping representations, activities that give S more experience connecting the population to the results of a sample may be needed. When S is able to successfully reason about expected values from known populations, S may be better able to gain more benefit from the ESD. In this case, activities such as the Betting King game in Schnell (2018) could form as both a useful intervention and an assessment of S's understanding of connections between population and sample.

5.1.3 Identifying Entities

S's attention to entities demonstrated facility with jumping between the *case* level and the *Set* level throughout the interviews, including in the pre-questions, and this attention to those two levels appeared to be a key support for their attention to the inter-level mechanism of swamping. Right from the Hospital problem in the first interview, S attended to the impact of individual boys born on the percentages of boys, and this appeared to be a powerful foundation for their reasoning throughout all the activities. Much of S's changes in reasoning throughout the interviews was not on the entities that they noticed, but on the properties and activities.

What is perhaps most noteworthy is how much S reasoned about individual samples even when presented with ESDs and TSDs. When asked about why an ESD was a certain way, S would explain by reference to what was likely or unlikely in a single sample. As noted previously, S also attended to the sample size plot instead of the ESD whenever the option was given to them, and their reasoning tended to be much richer in the context of the sample size plot than in the context of the ESD. It seems that, overall, the sample size plot was actually a support for their reasoning because of the connections between levels that it made visible, whereas the ESD was a somewhat opaque representation that might be able to be explained by representations that S had more connection to such as by reasoning about a single sample. S basically never referred to the TSD outside of the building blocks activity, although there were hints of permutation-based reasoning in the post-test. It may well be that S did not have a sufficiently abstract sense of the sample mean in order to successfully process the ESD and TSD. This lack of abstraction may have been partially helpful because it allowed S to avoid treating the mean like a "black box" and therefore to attend closely to swamping, unlike participants who overgeneralize proportional reasoning and thus ignore the effect of sample size. However, a lack of abstraction would also make reasoning about distributions of means such as the ESD and TSD quite challenging.

This lack of traction with the ESD is problematic, given that ESDs are a fundamental basis of the CATALST course and the principle vehicle for teaching inference (Zieffler & Catalysts for Change, 2017). S clearly had procedural knowledge about how to calculate a *p*-value from the ESD that they demonstrated during the Exam Preparation post-question, but had no way of reasoning about what the ESD might look like at different sample sizes. Although S did reason successfully with ESDs in several contexts through reference to individual samples, an inability to use ESDs intrinsically as a tool for reasoning and to engage with their properties may be a hindrance to successful statistical reasoning. In particular, S drew very little connections between the TSD and the ESD to reason about heaping or the future shape of the distribution, or to connect to the population mean. In the terms of Chi (2013), this seemed like a case where S's difficulty came from a *missing schema*: S did not have available categories of anything that functioned like a TSD or ESD, and thus was not able to use it fully as a reasoning tool or to develop more understanding of heaping. A lot of groundwork would need to be laid for S to develop understanding of the sampling distribution as a representation that could support understanding of sampling variability.

Despite the apparent missing schema, S did have momentary abilities to identify different permutations when dealing with a 0-1 variable. After viewing a plot that showed the different possible permutations that made up each outcome in the ESD, S reasoned effectively with the “little stacks” (3.1613–1614) of permutations, recognizing that certain means were more likely because they had more permutations that comprised them. However, this understanding was fragile, and permutations no longer were an entity when viewing analogous representations of a 0-1-1 variable. S’s familiarity with the 0-1 situation may have supported them to notice more subtle properties, whereas without even a clear sense of center in the 0-1-1 situation S was unable to note permutations.

5.1.4 Identifying Properties & Activities of Entities

Although Russ et al. (2008) gave “properties” and “activities” separate codes, and the present coding system also distinguished between them, they are here discussed together because properties and activities—both of entities and of entity properties—in practice are more features of the representations than they are features of the underlying process of the Empirical Law of Large Numbers. For instance, one of the most common activities was the increase in the size of a sample, since many of the “growing a sample” activities involved gradually increasing the sample size, but this is hard to substantively differentiate from S’s comparisons of large and small samples, since the “action” is just a way of representing different sample sizes. Therefore this section will treat properties and actions as synonymous.

S’s greatest confusion in the pre- and post-questions appeared to be around the property of *Set size*, or number of elements. S was eager to apply swamping reasoning to relate individual cases to the size of various sets, which only sometimes led to normative reasoning. In the Hospital and Post Office problem in the pre-questions, S conflated ESD size with sample size, noting that the smaller sample size would be expected to vary more, but then double-applying this reasoning that because there were a large number of days that the small and large sample sizes would then even out again and thus that they would be about the same. In the Post Office Simulation, S even applied this reasoning to their estimate of a region of a population distribution dotplot, revising their estimate of 25% down to 15% because “there’s a lot more people” (1.1124) and so each individual case would contribute less to the total, which therefore meant that an area that looked like 25% of the distribution was actually less. S at several points brought up an analogy to how someone with many $1s in their wallet would probably not notice if they lost one, whereas someone with much fewer would in fact notice. S’s reasoning corresponds to the well-known numerical cognition regularity of the distance and size effects (Moyer & Landauer, 1967), whereby the same distance ($1) takes longer to compare for larger numbers ($999 and $1000) as compared to smaller numbers ($9 and $10). Recent research by Obrecht (2019) found that sensitivity to sample size followed a curvilinear pattern, which is consistent with the predictions of sample size sensitivity operating on a power law, similar to psychophysical phenomenon. This relationship supports S in some complex reasoning about the sampling variability mechanism, but S overapplies it in the case of ESD size and the estimation task.

S almost appeared to be on the cusp of being able to differentiate these two levels of size in the post-interview when they correctly answered the Hospital problem. They wavered between choosing the smaller hospital, then changed their thinking and thought they would be about the same “because I forgot that was over a period of one year” (5.682–683). However, they then noted that the smaller hospital would vary more “looking at it individually” (5.693)—they appeared to be able to see that the problem was really asking about variability in days. If this was their understanding, however, it was still quite fragile, since they went on to conclude that the larger and smaller post offices would have about the same number of days with high extreme averages, noting that in “a week, the percentages might be drastically different, but if you look at the specific days I feel like it’d probably be pretty equivalent” (5.1323–1327). They appeared to be aware in this level that size was important on some levels but not on others, but did not have consistency on what those levels were.

S’s confusion regarding set size’s place in the mechanism appears to be a case of viewing size as a *special agent* in a *direct* (i.e. narrative) *schema*, as opposed to an emergent process (Chi et al., 2012). In a direct schema, an agent with special status directly causes an outcome, similar to the protagonist in a drama. However, as an emergent process, it is the collective summing of many independent agents that produces the effect of sample size. With S’s attention on “size” as an agent, it is difficult for them to distinguish when it would be associated with less sampling variability (sample size) vs. when it would be orthogonal to sampling variability (ESD size). In contrast, if S’s focus is on the collective summing mechanism, whereby it is the process of taking the mean or percentage that causes the emergent Empirical Law of Large Numbers, S would be able to see that sample size is relevant because it is involved in that collective summing, whereas ESD size is not.

Another explanation, which may be happening simultaneously, is simply a confusion of individual samples and the empirical sampling distribution. Well et al. (1990) noted that some participants repeated back the Hospital and Post Office Problems as if they were talking about a single sample, rather than about many samples, and S reflected later that they had originally been thinking about the whole year as a single big sample, rather than looking at each day. It may also be that the term “day” does not activate S’s schema of a “sample”, and so the mapping between the situation and sampling is not made clear to them and thus they may not view the days as individual samples.

Lexical ambiguity, the confusion between everyday usages of a word and the specialized way they are used in a domain (Kaplan, Fisher, & Rogness, 2009), appeared to play a role in S’s conflation as well. S used the word “trial”, which in the CATALST curriculum would indicate a single element of the ESD (Zieffler & Catalysts for Change, 2017), variously to refer to adding a single case onto a sample or to the sampling distribution, and they fluidly moved between using the number of trials to refer to sample size and ESD size. When examining an ESD during the Mystery Mean, S noted that they needed more “trials”; they noted that the trials should be 500 “for, like, collecting the statistic” (3.459–460) which is where one specifies ESD size in TinkerPlots™, but that the trials should be 200 “for, like, the draw on the original sampler” (3.460–461), an apparent reference to where one specifies sample size although the correct TinkerPlots™ field for that particular problem was actually “Repeat”. Although S briefly and correctly distinguished between their two meanings of “trials”, they frequently conflated them elsewhere and more clearly differentiated vocabulary might have supported them better. S sometimes seemed to cross both meanings in the same thought, leading to chimerical reasoning such as that “each point will be larger” (3.527–528) in the ESD when increasing the sample size.

S also seemed to struggle with lexical ambiguity between the words “varies” and “variability”. Variability had at least three meanings for S: range, multimodality, and number of possible means. When S used variability to mean range, this led to relatively normative reasoning because in most of the scenarios given here, range was a reasonable measure of variability that was consistent with what was taught regarding variability in the CATALST curriculum (Garfield et al., 2012; Zieffler & Catalysts for Change, 2017). However, S moved between variability-as-range and variability-as-multimodality when reasoning about Mystery Machine #2, without a clear sense of the distinction between the two. S also invoked variation when they were talking about the number of possible values two times—both times correcting themselves and noting that they were really talking about the number of possible means, but the concept of variation seemed to at least interfere with their clear description of the phenomenon of the number of possible means. S’s attention to possible means is discussed more in the next section.

5.1.5 Identifying Organization of Entities

Russ et al. (2008) defines the organization of entities as “how the entities are spatially organized, where they are located, and how they are structured” (p. 513). These abstract mechanisms did not have a spatial organization or location, but there were several structural relationships between entities and properties that S attended to. Here, the term structural relationships refers to basic non-stochastic mathematical relationships such as when S noted that they would divide the total by the sample size in order to get the mean (e.g. 1.944–953).

At several points early in the interviews, S noted the correspondence between sample size and *time*. This occurred even before being introduced to the sample size plot on the Batting Average pre-question, where S repeatedly emphasized that earlier in the season, there would be a lower sample size than later in the season, and that structural relationship was the only part they wrote in their written explanation. When introduced to the sample size plot in the second interview, S again repeatedly emphasized the connection between sample size increasing, time increasing, and the sample size plot "always, like, going up" (2.315). S's attention to this correspondence between sample size and time may have played into their confusion during the Hospital and Post Office problems, which both involved "over the course of a year" but without the sample size increasing, since each day was treated separately. The cue of time may have led S to think of these as accumulating a big sample over a year.

S also paid a lot of attention to the possible values that means or percentages could take on at different sample size, which they termed "slots" (2.200). S first attended to slots in Growing Sample Proportions: Physical Simulation, noting that at sample size 10 the possible percentages were every 10%. But it was when exploring ESDs and TSDs that S really spent more time examining slots, and in fact slots seemed to be more salient than other features (such as decreasing variability). Although S briefly attended to permutations correctly in Growing More Means and Growing Possibilities with a 0-1 variable, calling them “combinations”, when moving to a 0-1-1 variable S used the same word to actually indicate slots (which also corresponded to the mathematical meaning of combination). When asked about what they saw changing in the ESD as sample size grew during Growing Many Means, S repeatedly only responded that the number of slots was increasing. The Growing Certain sequence of activities did not particularly emphasize slots. Given the salience of slots, however, S may have benefited from more explicit exploration and attention to slots to understand how slots relate to the mechanism of Empirical Law of Large Numbers.

5.1.6 Chaining: Backward and Forwards

S appeared to come into the study with a strong sense of swamping, exhibiting strong chaining about the differential impact of cases on the percentages and mean during the pre-questions. However, as S developed more fine-grained descriptions of the target phenomenon and identified various aspects of the mechanism, documented above, S’s chaining about swamping gained more richness and subtlety. Moreover, as S was exposed to a variety of representations and contexts, their notion of swamping was complemented by other pieces of the mechanism of sampling variability.

S’s chaining about swamping in the pre-questions had little connection to random sampling, focusing only on the impact of cases at different sample sizes. S would talk about how a case in a small sample would have a greater impact than a case in a large sample, and then mention the long-run expected trend when this was clear to them, but there was not much of a connection between the swamping and the long-run trend. It may be that many of the textual problems did not elicit S’s thinking about samples, since problems like Hospital are not explicitly about samples. However, S displayed similar reasoning in the Post Office Simulation even when the sampling nature of the problem was made quite explicit. When reasoning with a known bell-shaped population in that activity, S was able to account for the long-run tendency by a combination of swamping and the population mode. Since most people will be at the mode, a single outlying value will *both* affect the sample mean very little at a large sample size *and* the sample mean will be near that population mode, which in the Post Office Simulation also happened to be the population mean. S invoked similar “swamping + mode” reasoning in the Cat Factory 1 activity, which also involved a bell-shaped known population distribution.

They then applied this same logic to their Cat Factory 2 sampler, which was a uniform distribution—which meant that the mean would never settle down, since all values were the mode. When this was disconfirmed, S displayed a hint of reasoning beyond the population mode as to why there would be a given long-run mean, noting that if there were both high and low numbers in a large sample, the mean would be in the “middle of that” (2.1298). Because of the cognitive conflict engendered when S realized that their existing sense of mechanism was inadequate to capture the phenomenon, S built onto their existing understanding of swamping and recognized the operation of *balancing*, the fact that large and small values would tend to cancel each other out (Well et al., 1990). Balancing accounted for why the sample mean would settle in the middle of a uniform distribution. Note that S never identified the population mean in this situation or the 0-1-1 situation, but still correctly accounted for it through the mechanism of balancing.

As noted in Chapter 3, balancing was not an intentional target of the activities, but it became an important resource for S. Unexpectedly, S's benefit from the supposedly heaping-focused activities in the fourth interview appeared to be largely due to how the activities supported their attention to and understanding of balancing. In the building blocks activity, S attended to how there were more permutations in the middle; when S transitioned to viewing the sample size plot, they then observed that "each time you'll probably get a mixture of black and white" (4.729–730). This new idea of "mixture" was consistent with their prior statements about permutations, but also built on their prior reasoning in Cat Factory 2 about how the mean would be in the middle because of the combination of high and low numbers. Then, in the post-questions, S built more on this concept of mixture for the Geology problem, noting that with a large sample they might have many measurements that were high, but they might also have many measurements that would be low and thus balance out, and that thus a smaller sample doesn't give a "good, um, mixture" (5.98). Seeing the different permutations piling up in the middle in the heaping plots may have supported S developing this concept that mixtures may help a sample be more representative, and that a large sample will be more likely to provide a good mixture.

Balancing appeared to play a role in S’s more generally thinking about long-run averages ending up in the “middle” even when the concept of mixture was less clear from their answers. Although S answered the Batting Average in a similar way on the post-test than on the pre-test, on the post-test S referred to how individual games may go better or worse, but the average would be “somewhere in the middle between your good and bad, it’s not going to be so much more towards, like, one way or the other” (5.961–964), suggesting an increased sense of the moderating effect of the average, whereas in the pre-questions their emphasis was simply on how they’d expect the average to decrease from a high extreme because “the baseballs are very small, so it’s very difficult to hit all of them” (1.663–664). Tying together the lesson they learned from Cat Factory 2 when asked to compare that representation with the Geology problem, S noted that even if all the probabilities above and below the true average in the Geology problem were the same, that “the average is gonna be somewhere in between whatever two, like, ways that it goes” (5.1404–1406).

S’s heaping reasoning was quite fragile and only occurred in a couple settings with maximal representational support and when examining small sample sizes of a 0-1 variable, where S had some intuitions about the population process. Outside of those settings, heaping was only recognizable by its byproducts, of which balancing was the most supportive of normative reasoning, as noted above. The other two byproducts could be termed *order irrelevance* and *increasing slots*. Order irrelevance was when S saw that multiple different orders could give the same mean. This supported S's understanding of heaping when examining the building-blocks permutations plots of the 0-1 variable, when S recognized that there would be more possible orders for certain means, making them more probable. At that time, S also noted that the order "won't affect the mean differently" (4.257–258), and it was this type of reasoning that seemed to outlast their sophisticated but fragile reasoning about the mean.

This order irrelevance reasoning came back in the Bottle Factory post-question, where S noted that having 10 perfect bottles at the beginning of the day "really doesn't matter" (5.207) for predicting whether there will be more bottles at the end of the day. S explicitly connected this with the 0-1 building blocks activity, noting that two blacks and two whites has the same order as two whites and two blacks and so "it doesn't really matter" (5.191) because there are still "the same amount of, like, white and black" (5.192–193). S again raised this issue in a similar way when comparing the permutations plot of a 0-1 variable with the Geology problem. Although this observation was essentially correct—different permutations of values can indeed equal the same mean—the connections with how certain means became more likely because of greater permutations being available did not seem to stay with S.

S's attention to the increase in the number of slots, as noted above, seemed to draw their attention away from other features of the ESD that might have supported them in reasoning about the mechanism. However, there is a connection between swamping and the number of slots that S's reasoning hinted at in several places: The increase in the density of slots causes there to be smaller and smaller possible movements in the mean, which is the phenomenon of swamping. S showed faint signs of this reasoning when doing the physical simulation for Growing Sample Proportions, when immediately after noting the decreasing changes in the proportion as sample size grew, they also noted that at sample size 10 that the slots would be at intervals of 10%. In close succession, S was attending to the size of the changes in the proportion, and the possible proportions themselves. After paying extensive attention to the increases in slots in the various heaping-focused activities, when S was interacting with both the sample size plot and the ESD in Growing Many Means, they were able to more explicitly link swamping and the increase in slots. S notes that "it's not moving quite as much back and forth" (4.1329–1330) because "it just kind of keeps on getting smaller—like, the spots, like, in between those numbers" (4.1347–1349).

5.2 Supporting Mechanistic Reasoning about the Empirical Law of Large Numbers

Although there are many prior studies of students' understanding of sampling variability and many other statistical concepts, this study was unique in statistics education research in that it focused on students' mechanistic reasoning specifically. The richness of the reasoning that this focus exposed suggests that mechanistic reasoning could be a fruitful area of study more generally in statistics education. S clearly had strong senses of mechanism before any intervention, and S's mechanistic reasoning was both malleable and influential on their overall statistical reasoning. Statistics educators increasingly value students' conceptual understanding of statistics beyond just understanding how to do rote procedures (e.g., delMas, Garfield, Ooms, & Chance, 2007). Supporting students in understanding mechanisms provides a way for students to understand *why* and *how* statistics works to support conceptual growth by forming inter-level causal explanations for students to update their understanding (Chi et al., 2012; Kendeou, Smith, & O'Brien, 2013). This can be extended to important statistical topics such as students' understanding of the conclusions relative to random sampling versus random assignment (Fry, 2017), links between the Poisson and exponential distributions (Budgett & Pfannkuch, 2010), the inference of comparing groups (Wild, Pfannkuch, Regan, & Horton, 2011), and many other topics. Mechanistic reasoning is valued in science education generally (Russ et al., 2008), and there is no reason for statistics to be an exception.

Mechanistic reasoning may be particularly crucial for exploring the Empirical Law of Large Numbers because this principle is the justification for the connection between the empirical and "true" probability (Schnell, 2018). Prior studies have tried to teach this fundamental connection via a variety of experiences, but there can be major conceptual challenges in seeing the experimental probability as an estimate of the true probability (Konold et al., 2011). Mechanistic reasoning provides a different possible connection by showing how and why the empirical probability converges to the true probability via swamping, balancing, and heaping, rather than merely demonstrating it in examples, or by proving the mathematical law of large numbers or central limit theorems which are, as Sedlmeier and Gigerenzer (1997) have pointed out, asymptotic theorems which do not actually explain what is happening at finite sample sizes.

Mechanistic reasoning at some level about the Empirical Law of Large Numbers was feasible and displayed on many occasions by all five participants (and the five pilot participants). Moreover, the detailed analysis of S's responses suggested that this mechanistic reasoning can be shaped and supported through representations and prompts, which seemed to lead to at least short-term changes in reasoning. This rich mechanistic reasoning, and the manner in which even the pre-questions elicited different types of mechanistic reasoning, imply that mechanistic reasoning may not need to be considered a totally new target of instruction. The participants, as highlighted in the case of S, came to the interviews with complex mechanistic reasoning already. S drew on rich and relevant former experiences of variability in their analogies, drawing on experiences of flipping a coin, noticing how the price of eating out made a much bigger impact on them as a college student than on a richer person, and recalling the uncertainty of whether there would be blue Scooby Doo fruit snacks in their bag of candy, and they naturally had a sense of mechanism (diSessa, 1993) that they drew upon for explaining these experiences and the ones they encountered during the interview. S gave a non-normative answer to the Hospital and Post Office problems not because they had a "representativeness heuristic" (Kahneman & Tversky, 1982) that led them to ignore sample size, but because they knew sample size was important but had trouble mapping their understanding and experience onto the mechanism of the empirical sampling distribution. The role of an instructor teaching S would be to understand their sense of mechanism and to provide experiences that helped them to refine, connect, develop, and extend their mechanistic reasoning about sampling variability to a broader range of situations and experiences (Wagner, 2006).

Growing Certain had mixed success in supporting S's mechanistic reasoning. The growing a sample paradigm (Bakker, 2004), and the sample size plot in TinkerPlots™ in particular, seemed to provide a valuable tool for S viewing the impact of sample size as a process. This allowed them to more clearly articulate the phenomenon and identify key mechanistic elements, centered on swamping reasoning, to support quite rich chains of reasoning and allowing them to play the mechanism forwards and backwards. However, S dealt with the limits of swamping reasoning—its inadequacy for describing long-run convergence—in a different way than expected. Instead of complementing swamping with heaping, S instead complemented swamping with balancing, and seemed to draw on the heaping representations to support their use of balancing in the later interviews. Balancing was both more accessible, without the requirement for attending to differential rates of change in a theoretical sampling distribution required by heaping, and it satisfied the goal of accounting for the location of the sample size plot in the situations encountered in these interviews. Future research is needed to see how balancing could be more directly supported as a way of complementing swamping to understand the mechanism of sampling variability.

The overall failure of the development of heaping reasoning is likely due to a combination of S's numeracy level, the inherent complexity of heaping, and the lack of a bridge between the change of a single sample and examining many samples either in

Growing Certain or in S’s prior educational experiences. Besides for S’s difficulty mapping their mechanistic reasoning of sampling variability onto an empirical sampling distribution, S seemed to lack a sense of the empirical sampling distribution as an expression of the possibilities for what could happen for a given sample mean—the one-to-many connection was missing, leading S to fall back on to the single sample size plot for making future predictions even when an ESD was available that answered the question at hand. Another advantage of the sample size plot was that it directly represented the changes in the mean caused by additional cases in the sample, whereas these were abstracted away and hidden in the ESD and even in the permutations plots made with building blocks. This is despite the fact that S reasoned about sampling variability exclusively using ESDs in the CATALST course (Zieffler & Catalysts for Change, 2017).

How could a bridge be built between the changes in the sample and the shape of the resulting ESD and TSD? There is evidence in S’s responses that going from one mean growing to a whole distribution of means is just too big a leap. S had a glimpse of more reasoning about multiple means in the Growing Many Means activity when they viewed just a handful of means growing simultaneously, with a sample size plot of one of those means below (Figure 4.51). S may be able to find more traction in understanding ESDs, and how they relate to the underlying samples, if they actually have a clear mechanistic understanding of the relationships of the ESDs to the underlying samples. This appeared to require not just growing a single sample and then looking at many samples, but instead to show the movement of the mean, and then the movement of a handful of means simultaneously, to provide a more trackable representation with clearer relationships to the underlying sample size plot. Overlaying multiple sample size plots, or incorporating animations that switch between different samples of the same size to show the differences in those sample size plots (Hullman et al., 2015), may be ways of supporting this bridge. For instance, three sample means could be shown growing simultaneously, both in a sample size plot and in a dotplot that shows just those means similar to the Growing Many Means; once students have reasoned about this, it could be extended to more means simultaneously to build toward an understanding of the ESD. Although the exact form of the best representation is not known, some kind of representation and activity that supports students in making the connections between the *changes* in an individual sample and the *distribution* of many samples clearly seems called for. This seems like an important step to take before exploring heaping.

Another possible step forward for teaching heaping is to make the connection between the sample size plot and the theoretical sampling distribution more explicit by presenting the theoretical distribution in a similar way to the building blocks activity, but to make it more clear early on that each permutation of values is a *possible sample size plot*, and that heaping occurs because there are more possible sample size plots that land near the population mean. This could be done by having linked representations of the permutations in a building-blocks-like representation, and their corresponding sample size plots. This might allow students to see the full possible range of sample size plots and to see the connection of a particular plot to its corresponding permutation. This was attempted in a limited way in the Growing Possibilities activities by having students draw the sample size plots that corresponded to certain permutations, but the connection could be made more clear early on. For instance, connections between levels could be made clearer by having students draw all the sample size plots that start in the same way as a sample size plot on the previous level, and only then create the permutation using the building blocks and placing them on the line to help them see the mechanism's connections.

Finally, heaping should move beyond binomial situations by using software to represent the different possibilities for continuous populations. The Cat Factory activity was quite successful in prompting rich reasoning about different population shapes because S could manipulate the population directly, observe the results, and infer the mechanistic effect of the population. Allowing S to specify two different populations that they would expect to have differing TSD shapes, with the ability to view underlying sample size plots as well, could prompt richer reasoning about heaping and more ability to draw on heaping when reasoning about ESDs.

5.3 Mechanistic Coding to Explore Student Understanding of Sample Size

Identifying and distinguishing the abstract entities, properties, and activities in sampling variability presented some methodological challenges for the present study. S used nearly all levels of the Russ et al. (2008) hierarchy even when just answering the Hospital problem on the pre-interview, so that coding alone was not informative. It was more informative to code what the entities, activities, and properties actually were, which helped explore both the richness and the conflations in S's reasoning. More specific coding allowed this study to track how entities, activities, and properties changed across the different activities.

The complex coding scheme that resulted, however, did not always have clear boundaries between entities, properties, and actions, and often there was ambiguity in coding the relationships between entities and properties. This coding was therefore not well-specified enough to be replicable and reliable. Additionally, the codes were assigned by just one person, which is why the codes were reported at such a general level in Chapter 4. Future studies attempting to use a mechanistic grammar to characterize student thinking needs to make some additional decisions to specify boundaries between codes, especially regarding relationships, in order to enable reliability and replicability. Such an effort most likely requires a team that includes experienced coders, with regular lab meetings to make decisions to define the codebook in a pilot phase. This team would create multiple independent codings of the same sections in order to evaluate inter-rater reliability, and then to return to meetings to work out consensus. Such group-coding exercises were beyond the scale of this dissertation.

A particular challenge may emerge when attempting to apply the coding scheme across participants, and compare codes across participants. Because this study examined only one participant in detail, coding could keep relatively close to S's words; for instance, S used the word "slots" in a particular way and this became its own entity, but others may group concepts in subtly different ways while using the same words—or group similarly and use different words, and which one they are doing may not be clear from the data.

Again, multiple coders and in-depth discussion of the coding system would probably be the best way to work out these difficulties.

5.4 Teaching Implications

This close analysis of one students' mechanistic reasoning about sampling variability has potential implications beyond instruction specifically regarding the mechanisms of the Empirical Law of Large Numbers. S's ability to reason much more successfully with sample size plots than with ESDs suggests that their ability to work with ESDs conceptually is much more limited than would be useful for understanding the dynamics of inference, despite the centrality that ESDs play in the CATALST curriculum. If students do not have the ability to interpret the ESD as a probability distribution for the means of samples, then simulation-based inference is simply substituting the opacity of asymptotic formulas for the opacity of the ESD. If we wish for students to understand the dynamics of inference, such as to be able to predict how we would expect the *p*-value to change when the sample size increases as in the Exam Preparation Strategies problem, then more care should be put in to how students understand the ESD and how the ESD emerges from the growth of many samples. The strategies outlined above regarding building up the ESD by first growing a small number of means at once with simultaneous sample size plots could be useful not just for understanding the mechanism of the Empirical Law of Large Numbers, but more generally for understanding what the ESD actually represents and how it relates to what *could* happen in an individual sample. This relationship is central to statistical inference.

Another potential implication for teaching is providing clearer representational, notational, and lexical distinctions between entities and properties that S often conflated both verbally and conceptually. S applied “variability” and “spread” to multimodality and the number of distinct possible values, and “trials” both to individual cases and to individual samples. Teachers should take care in clearly differentiating these concepts, especially regarding the different levels and entities of sampling variability, and be thoughtful about lexical ambiguity between everyday and specialized usage of terms (Kaplan et al., 2009).

5.5 Limitations & Opportunities for Future Research

An obvious irony of this study was its very small sample size of one student. All five study participants, and the five pilot participants, had their own idiosyncrasies in reasoning; as reported elsewhere, one of the pilot participants was particularly resistant to swamping reasoning even after extensive support (Brown, 2018), whereas S employed it immediately on the first pre-question. Therefore the comments here about supporting other types of reasoning via the sample size plot and swamping should be taken with caution. As a next step, the other four study participants could be analyzed to a similar level of detail to see how well these conclusions generalize across the sample. Moreover, more information about the participants themselves should be gathered and compared for future research about mechanistic reasoning about the Empirical Law of Large Numbers. S’s numeracy had some apparent effects on their responses, and there is evidence that numeracy correlates with greater sensitivity to sample size (Obrecht, 2019). S was also not as sensitive to contradictions in their reasoning, so measures such as the Cognitive Reflection Task (Frederick, 2005) could help explore the role that dual-processing might be playing in mechanistic reasoning about sample size (see also Lem, 2015). And given S’s use of prior experiences of a sample size such as their perceptions of wealthy and poorer people’s noticing of money leaving their wallet—or another participant’s extensive use of genetics analogies that were sparked by their mixed-race heritage—background and demographic factors would be important to consider in future work.

It should also be noted that these students were in the CATALST course (Zieffler & Catalysts for Change, 2017), which is by no means a typical introductory statistics curriculum (Garfield et al., 2012). Growing Certain was very specifically geared towards this curriculum, building directly off of the activities and representations used in this course, and substantial support would need to be given for students in other curricula to participate and engage with these tasks and representations. If anything, S’s difficulties with ESDs might be multiplied many times for students who did not have the carefully structured series of activities in CATALST to prepare them to set up TinkerPlots™ models and interpret dotplots of means and percentages.

Moreover, this study only makes clear the mechanistic reasoning that S displayed, but does not clearly identify what produced that reasoning. There were no controlled manipulations as a part of these activities that would be able to separate precisely what contributed to S’s reasoning. The unusual social situation of talking in detail for six hours about sample size with an interested interviewer who validated their reasoning no matter what they said, and who asked many questions about why they thought certain things were true, may have produced rich mechanistic reasoning even without any of the Growing Certain activities. S seemed to enjoy and be encouraged to dive deeper into their thinking with the interviewer's positive attention, whereas other participants felt it was awkward to be constantly validated and not told what the "right answer" was.

And what is the "right answer"? The normative expectations for mechanistic reasoning about the Empirical Law of Large Numbers deserve further mathematical study and communication to the statistics education community. It is possible that mathematicians and theoretical statisticians have already developed solutions to these problems, but if so these results are not represented in the statistics education literature. Sedlmeier and Gigerenzer (1997) characterize the Empirical Law of Large Numbers as a "common-sensical intuition" (p. 35) that mathematically only has "partial justifications" (p. 48) in the formula for the variance of the sample mean, in Chebychev's inequality, and the central limit theorem. However, there are likely stronger and mathematical justifications that can be built for many typical applications of the Empirical Law of Large Numbers even beyond the uncontroversial binomial examples explored in the Growing More Means and Growing Possibilities activities. As discussed in Chapter 2, swamping is a mathematically uncontroversial mechanism that follows immediately from the formula for the mean, but it only captures a small part of the implications of the Empirical Law of Large Numbers since it does not explain convergence to a population mean. More detailed explorations of the mathematical phenomenon underlying balancing and heaping, on the other hand, may provide a richer normative account of the mechanism, and how and when it applies. Having this normative standard more clearly specified could help statistics educators design curricula that would support development of understanding the mechanism, and more clearly evaluate students' understanding of the mechanism. In particular, this normative standard would provide guidance for designing instruction to guide students in drawing connections between a true probability and the empirical outcomes observed, which is a fundamental issue in understanding probability and statistics (Konold et al., 2011; Schnell, 2018).

This study took the first small steps in exploring students' mechanistic reasoning about the Empirical Law of Large Numbers. Given the richness of the data produced, and the potential implications of mechanistic reasoning for understanding statistical phenomena and developing lasting conceptual understanding, there are many rich avenues for further research that could build on the strengths and weaknesses of the present work.

References

Abrahamson, D. (2006). The Shape of Things to Come: The Computational Pictograph as a Bridge From Combinatorial Space to Outcome Distribution. *International Journal of Computers for Mathematical Learning*, *11*(1), 137–146. https://doi.org/10.1007/s10758-006-9102-y

Amsel, E., Klaczynski, P. A., Johnston, A., Bench, S., Close, J., Sadler, E., & Walker, R. (2008). A dual-process account of the development of scientific reasoning: The nature and development of metacognitive intercession skills. *Scientific Reasoning -- Where Are We Now? Sodian and Bullock SI*, *23*(4), 452–471. https://doi.org/10.1016/j.cogdev.2008.09.002

Bakker, A. (2004, May 13). Design research in statistics education : on symbolizing and computer tools [Dissertation]. Retrieved February 20, 2017, from http://dspace.library.uu.nl/handle/1874/893

Bar-Hillel, M. (1979). The role of sample size in sample evaluation. *Organizational Behavior and Human Performance*, *24*(2), 245–257. http://dx.doi.org/10.1016/0030-5073(79)90028-X

Bar-Hillel, M. (1982). Studies of representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.

Ben-Zvi, D. (2006). Scaffolding students' informal inference and argumentation. *Proceedings of the Seventh International Conference on Teaching Statistics*.

Brown, E. (2015). *Sample Size Reasoning of Students in a Randomization-Based Introductory Statistics Course* (Unpublished Master's thesis). University of Minnesota, Twin Cities.

Brown, E. (2018). Developing students' causal understanding of sampling variability: A design research study. *Proceedings of the 10th International Conference on Teaching Statistics (ICOTS-10)*. Presented at the Kyoto, Japan. Kyoto, Japan: International Statistical Institute.

Budgett, S., & Pfannkuch, M. (2010). Assessing students' statistical literacy. *Assessment Methods in Statistical Education: An International Perspective*, *19*, 103.

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). Retrieved from http://dx.doi.org/10.1007/1-4020-2278-6_13

Chi, M. T. H. (2013). *Two Kinds and Four Sub-Types of Misconceived Knowledge, Ways to Change it, and the Learning Outcomes*. Retrieved from https://www.routledgehandbooks.com/doi/10.4324/9780203154472.ch3

Chi, M. T. H., Roscoe, R. D., Slotta, J. D., Roy, M., & Chase, C. C. (2012). Misconceived Causal Explanations for Emergent Processes. *Cognitive Science*, *36*(1), 1–61. https://doi.org/10.1111/j.1551-6709.2011.01207.x

Cobb, G. W. (2007). The Introductory Statistics Course: A Ptolemaic Curriculum? *Technology Innovations in Statistics Education*, *1*(1). Retrieved from http://www.escholarship.org/uc/item/6hb3k0nz

Common Core State Standards Initiative. (2010). *Common Core State Standards for Mathematics (CCSSM). Washington, DC: National Governors Association Center for Best Practices and the Council of Chief State School Officers*.

Creswell, J. W. (2012). *Qualitative inquiry and research design: Choosing among five approaches*. Sage.

DeCarli, E. (2012, July 6). Exploring the Law of Large Numbers with TinkerPlots. Retrieved June 4, 2018, from Sine of the Times website: http://www.sineofthetimes.org/exploring-the-law-of-large-numbers-with-tinkerplots/

delMas, R. (2002). Sampling SIM (Version 5.4). Retrieved from http://www.tc.umn.edu/~delma001/stat_tools/

delMas, R., Garfield, J., & Chance, B. (1999). A model of classroom research in action: Developing simulation activities to improve students' statistical reasoning. *Journal of Statistics Education*, *7*(3).

delMas, R., Garfield, J., & Chance, B. (2006). *Using assessment to study students' reasoning about sampling distributions*. Unpublished manuscript.

delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, *6*(2), 28–58.

delMas, R., & Liu, Y. (2005). Exploring students' conceptions of the standard deviation. *Statistics Education Research Journal*, *4*(1), 55–82.

Dinov, I. (2017). SOCR Data Dinov 020108 HeightsWeights - Socr. Retrieved June 4, 2018, from Statistics Online Computational Resource website: http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_Dinov_020108_HeightsWeights

diSessa, A. A. (1993). Toward an Epistemology of Physics. *Cognition and Instruction*, *10*(2/3), 105–225.

diSessa, A. A. (2007). An Interactional Analysis of Clinical Interviewing. *Cognition and Instruction*, *25*(4), 523–565. https://doi.org/10.1080/07370000701632413

Evans, J. S. B. T., & Dusoir, A. E. (1977). Proportionality and sample size as factors in intuitive statistical judgement. *Acta Psychologica*, *41*(2–3), 129–137. https://doi.org/10.1016/0001-6918(77)90030-0

Fiedler, K., Walther, E., & Nickel, S. (1999). The auto-verification of social hypotheses: Stereotyping and the power of sample size. *Journal of Personality and Social Psychology*, *77*(1), 5–18. https://doi.org/10.1037/0022-3514.77.1.5

Finzer, W., Erickson, T., & Binker, J. (2005). *Fathom: Dynamic statistics software*. Retrieved from http://fathom.concord.org

Fischbein, E., & Schnarch, D. (1997). The Evolution with Age of Probabilistic, Intuitively Based Misconceptions. *Journal for Research in Mathematics Education*, *28*(1), 96–105. https://doi.org/10.2307/749665

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology*, *18*(3), 253–292. https://doi.org/10.1016/0010-0285(86)90001-0

Fong, G. T., & Nisbett, R. E. (1991). Immediate and delayed transfer of training effects in statistical reasoning. *Journal of Experimental Psychology: General*, *120*(1), 34–45. https://doi.org/10.1037/0096-3445.120.1.34

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2005). *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A Pre-K-12 Curriculum Framework*. Retrieved from American Statistical Association website: http://www.amstat.org/education/gaise

Frederick, S. (2005). Cognitive Reflection and Decision Making. *The Journal of Economic Perspectives*, *19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Freudenthal, H. (1972). The 'empirical law of large numbers' or 'The stability of frequencies.' *Educational Studies in Mathematics*, *4*(4), 484–490. https://doi.org/10.1007/BF00567002

Fry, E. B. (2017). *Introductory Statistics Students' Conceptual Understanding of Study Design and Conclusions* (Ph.D., University of Minnesota). Retrieved from http://login.ezproxy.lib.umn.edu/login?url=https://search.proquest.com/docview/2025494905?accountid=14586

GAISE College Report ASA Revision Committee. (2016). *Guidelines for Assessment and Instruction in Statistics Education College Report 2016*. Retrieved from American Statistical Association website: http://www.amstat.org/education/gaise

Gal, I. O. (1989). *Which Group Is Better? The Development of Statistical Reasoning in Elementary School Children.* Retrieved from https://eric.ed.gov/?id=ED315270

Garfield, J., delMas, R., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM*, *44*(7), 883–898. https://doi.org/10.1007/s11858-012-0447-5

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review*, *103*(3), 592–596. https://doi.org/10.1037/0033-295X.103.3.592

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, *24*(3), 411–435. https://doi.org/10.1016/0010-0285(92)90013-R

Groth, R. E. (2005). An investigation of statistical thinking in two different contexts: Detecting a signal in a noisy process and determining a typical value. *The Journal of Mathematical Behavior*, *24*(2), 109–124. https://doi.org/10.1016/j.jmathb.2005.03.002

Hardiman, P. T., Well, A. D., & Pollatsek, A. (1984). Usefulness of a balance model in understanding the mean. *Journal of Educational Psychology*, *76*(5), 792–801. https://doi.org/10.1037/0022-0663.76.5.792

Hullman, J., Resnick, P., & Adar, E. (2015). Hypothetical Outcome Plots Outperform Error Bars and Violin Plots for Inferences about Reliability of Variable Ordering. *PLOS ONE*, *10*(11), e0142444. https://doi.org/10.1371/journal.pone.0142444

Ireland, S., & Watson, J. M. (2009). Building a Connection Between Experimental and Theorectical Aspects of Probability. *International Electronic Journal of Mathematics Education*, *4*(3), 339–370.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430–454. https://doi.org/10.1016/0010-0285(72)90016-3

Kahneman, D., & Tversky, A. (1982). On the study of statistical intuitions. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.

Kaplan, J. J., Fisher, D. G., & Rogness, N. T. (2009). Lexical Ambiguity in Statistics: What do Students Know about the Words Association, Average, Confidence, Random and Spread? *Journal of Statistics Education*, *17*(3), null. https://doi.org/10.1080/10691898.2009.11889535

Kendeou, P., & O'Brien, E. J. (2014). The Knowledge Revision Components (KReC) Framework: Processes and Mechanisms. In *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences* (p. 353). Retrieved from https://books.google.com/books?hl=en&lr=&id=TAFzBAAAQBAJ&oi=fnd&pg=PA353&dq=Knowledge+Revision+Components+framework&ots=vlQbt8OeKY&sig=ASKVU6tVSaYqY7ZlcmZH1lS-IRA

Kendeou, P., Smith, E. R., & O'Brien, E. J. (2013). Updating during reading comprehension: Why causality matters. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 854–865. https://doi.org/10.1037/a0029468

Konold, C., & Harradine, A. (2014). Contexts for Highlighting Signal and Noise. In T. Wassong, D. Frischemeier, P. R. Fischer, R. Hochmuth, & P. Bender (Eds.), *Mit Werkzeugen Mathematik und Stochastik lernen – Using Tools for Learning Mathematics and Statistics* (pp. 237–250). Retrieved from http://dx.doi.org/10.1007/978-3-658-03104-6_18

Konold, C., Harradine, A., & Kazak, S. (2007). Understanding Distributions by Modeling Them. *International Journal of Computers for Mathematical Learning*, *12*(3), 217–230. https://doi.org/10.1007/s10758-007-9123-1

Konold, C., & Kazak, S. (2008). Reconnecting Data and Chance. *Technology Innovations in Statistics Education*, *2*(1). Retrieved from http://www.escholarship.org/uc/item/38p7c94v

Konold, C., Madden, S., Pollatsek, A., Pfannkuch, M., Wild, C., Ziedins, I., … Kazak, S. (2011). Conceptual Challenges in Coordinating Theoretical and Data-centered Estimates of Probability. *Mathematical Thinking and Learning*, *13*(1–2), 68–86. https://doi.org/10.1080/10986065.2011.538299

Konold, C., & Miller, C. D. (2017). TinkerPlots: Dynamic data exploration (Version 2.3.1). Emeryville, CA: Learn Troop.

Konold, C., & Pollatsek, A. (2002). Data Analysis as the Search for Signals in Noisy Processes. *Journal for Research in Mathematics Education*, *33*(4), 259–289. https://doi.org/10.2307/749741

Leavy, A. M. (2006). Using data comparison to support a focus on distribution: Examining preservice teacher's understandings of distribution when engaged in statistical inquiry. *Statistics Education Research Journal*, *5*(2), 89–114.

Lee, H. S., Angotti, R. L., & Tarr, J. E. (2010). Making comparisons between observed data and expected outcomes: students' informal hypothesis testing with probability simulation tools. *Statistics Education Research Journal*, *9*(1), 68–96.

Lehrer, R., & Schauble, L. (2007). Contrasting emerging conceptions of distribution in contexts of error and natural variation. In M. C. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 149--176). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Lem, S. (2015). The intuitiveness of the law of large numbers. *ZDM*, *47*(5), 783–792. https://doi.org/10.1007/s11858-015-0676-5

Lem, S., Van Dooren, W., Gillard, E., & Verschaffel, L. (2011). Sample size neglect problems: A critical analysis. *Studia Psychologica*, *53*(2), 123–135.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about Mechanisms. *Philosophy of Science*, *67*(1), 1–25. https://doi.org/10.2307/188611

Marnich, M. A. (2008). *A Knowledge Structure for the Arithmetic Mean: Relationships Between Statistical Conceptualizations and Mathematical Concepts*. ProQuest.

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis : a methods sourcebook* (Third edition..). Thousand Oaks, Califorinia: Thousand Oaks, Califorinia : SAGE Publications, Inc.

Mokros, J., & Russell, S. J. (1995). Children's Concepts of Average and Representativeness. *Journal for Research in Mathematics Education*, *26*(1), 20–39. https://doi.org/10.2307/749226

Morgan, K. L., Lock, R. H., Lock, P. F., Lock, E. F., & Lock, D. F. (2014). StatKey: Online tools for bootstrap intervals and randomization tests. *Proceedings of the 9th International Conference on Teaching Statistics (ICOTS-9)*. Presented at the Flagstaff, Arizona, USA. Flagstaff, Arizona, USA: International Statistical Institute.

Moyer, R. S., & Landauer, T. K. (1967). Time required for Judgements of Numerical Inequality. *Nature*, *215*(5109), 1519–1520. https://doi.org/10.1038/2151519a0

Murray, J., Iding, M., Farris, H., & Revlin, R. (1987). Sample-size salience and statistical inference. *Bulletin of the Psychonomic Society*, *25*(5), 367–369. https://doi.org/10.3758/BF03330369

Nisbett, R. E., Fong, G. T., Lehman, D. R., & Cheng, P. W. (1987). Teaching Reasoning. *Science*, *238*(4827), 625–631.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review*, *90*(4), 339–363. https://doi.org/10.1037/0033-295X.90.4.339

Obrecht, N. A. (2019). Sample size weighting follows a curvilinear function. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(4), 614–626. https://doi.org/10.1037/xlm0000615

Obrecht, N. A., Chapman, G. B., & Suárez, M. T. (2010). Laypeople do use sample variance: The effect of embedding data in a variance-implying story. *Thinking & Reasoning*, *16*(1), 26–44. Retrieved from keh.

Obrecht, N. A., Chapman, G., & Gelman, R. (2007). Intuitive t tests: Lay use of statistical information. *Psychonomic Bulletin & Review*, *14*(6), 1147–1152. https://doi.org/10.3758/BF03193104

O'Dell, R. S. (2012). The Mean as Balance Point. *Mathematics Teaching in the Middle School*, *18*(3), 148–155. https://doi.org/10.5951/mathteacmiddscho.18.3.0148

Olson, C. L. (1976). Some apparent violations of the representativeness heuristic in human judgment. *Journal of Experimental Psychology: Human Perception and Performance*, *2*(4), 599–608. https://doi.org/10.1037/0096-1523.2.4.599

Parnafes, O., & diSessa, A. A. (2013). Microgenetic learning analysis: A methodology for studying knowledge in transition. *Human Development*, *56*(1), 5–37.

Piaget, J., & Inhelder, B. (1951). *La genèse de l'idée de hasard chez l'enfant. [The genesis of the idea of chance in the child.]*. 75006 Paris, France: Presses Universitaires de France.

Pollard, P., & Evans, J. S. B. T. (1983). The role of "representativeness" in statistical inference: A critical appraisal. In J. S. B. T. Evans (Ed.), *Thinking and reasoning: Psychological approaches* (pp. 107--134).

Pollatsek, A., Lima, S., & Well, A. D. (1981). Concept or computation: Students' understanding of the mean. *Educational Studies in Mathematics*, *12*(2), 191–204. https://doi.org/10.1007/BF00305621

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education*, *66*(2), 211–227. https://doi.org/10.1002/sce.3730660207

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.

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. Retrieved from https://www.R-project.org/

Reagan, R. T. (1989). Variations on a seminal demonstration of people's insensitivity to sample size. *Organizational Behavior and Human Decision Processes*, *43*(1), 52–57. https://doi.org/10.1016/0749-5978(89)90057-5

Russ, R. S., Scherr, R. E., Hammer, D., & Mikeska, J. (2008). Recognizing mechanistic reasoning in student scientific inquiry: A framework for discourse analysis developed from philosophy of science. *Science Education*, *92*(3), 499–525. https://doi.org/10.1002/sce.20264

Saldanha, L. A., & 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

Schnell, S. (2018). Integrating theoretical and empirical considerations: Young students'understanding of the empirical law of large numbers. *Proceedings of the 10th International Conference on Teaching Statistics (ICOTS-10)*. Presented at the Kyoto, Japan. Kyoto, Japan: International Statistical Institute.

Sedlmeier, P. (1998). The distribution matters: two types of sample-size tasks. *Journal of Behavioral Decision Making*, *11*(4), 281–301. https://doi.org/10.1002/(SICI)1099-0771(1998120)11:4<281::AID-BDM302>3.0.CO;2-U

Sedlmeier, P. (1999). *Improving statistical reasoning: Theoretical models and practical implications*. Mahwah, New Jersey: Lawrence Erlbaum.

Sedlmeier, P., & Gigerenzer, G. (1997). Intuitions about sample size: the empirical law of large numbers. *Journal of Behavioral Decision Making*, *10*(1), 33–51. https://doi.org/10.1002/(SICI)1099-0771(199703)10:1<33::AID-BDM244>3.0.CO;2-6

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology*, *8*(4), 481–520. https://doi.org/10.1016/0010-0285(76)90016-5

Siegler, R. S. (2007). Microgenetic Analyses of Learning. In *Handbook of Child Psychology*. https://doi.org/10.1002/9780470147658.chpsy0211

Siegler, R. S., & Crowley, K. (1991). The microgenetic method: A direct means for studying cognitive development. *American Psychologist*, *46*(6), 606–620. https://doi.org/10.1037/0003-066X.46.6.606

Stohl-Drier, H. (2000). The Probability Explorer: A research-based microworld to enhance children's intuitive understandings of chance and data. *Focus on Learning Problems in Mathematics*, *22*(3/4), 165–178.

Strauss, S., & Bichler, E. (1988). The Development of Children's Concepts of the Arithmetic Average. *Journal for Research in Mathematics Education*, *19*(1), 64–80. https://doi.org/10.2307/749111

Stroup, W. M., & Wilensky, U. (2014). On the Embedded Complementarity of Agent-Based and Aggregate Reasoning in Students' Developing Understanding of Dynamic Systems. *Technology, Knowledge and Learning*, *19*(1–2), 19–52. https://doi.org/10.1007/s10758-014-9218-4

Tintle, N., Topliff, K., Vanderstoep, J., Holmes, V.-L., & Swanson, T. (2012). Retention of Statistical Concepts in a Preliminary Randomization-Based Introductory Statistics Curriculum. *Statistics Education Research Journal*, *11*(1).

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105–110. https://doi.org/10.1037/h0031322

Utts, J. (2003). What Educated Citizens Should Know About Statistics and Probability. *The American Statistician*, *57*(2), 74–79. https://doi.org/10.1198/0003130031630

Vosniadou, S., & Skopeliti, I. (2014). Conceptual Change from the Framework Theory Side of the Fence. *Science & Education*, *23*(7), 1427–1445. https://doi.org/10.1007/s11191-013-9640-3

Wagner, J. F. (2006). Transfer in Pieces. *Cognition and Instruction*, *24*(1), 1–71. https://doi.org/10.1207/s1532690xci2401_1

Watson, J., & Moritz, J. (1998). The Beginning of Statistical Inference: Comparing two Data Sets. *Educational Studies in Mathematics*, *37*(2), 145–168. https://doi.org/10.1023/A:1003594832397

Well, A. D., Pollatsek, A., & Boyce, S. J. (1990). Understanding the effects of sample size on the variability of the mean. *Organizational Behavior and Human Decision Processes*, *47*(2), 289–312. http://dx.doi.org/10.1016/0749-5978(90)90040-G

Wild, C. J., Pfannkuch, M., Regan, M., & Horton, N. J. (2011). Towards more accessible conceptions of statistical inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *174*(2), 247–295.

Zieffler, A., & Catalysts for Change. (2017). *Statistical Thinking: A Simulation Approach to Modeling Uncertainty* (4.0). Minneapolis: Catalyst Press.

Appendix A: Correspondence with EPSY 3264 students

Appendix A1: Pilot recruitment first contact email

Subject: Research participation opportunity to improve statistics teaching

I'm Ethan Brown, a PhD candidate in Educational Psychology (advised by Associate Professor Robert delMas), and I'm researching new approaches to teaching statistics. I've developed some new activities similar to the ones you used in EPSY 3264: Basic and Applied Statistics last semester and was hoping you'd be willing to participate in a pilot study that consists of a series of interviews to explore your thinking about these new activities.

You are exactly who I need for the study: a student who has completed EPSY 3264. By participating, you'd help me to modify the activities and the data collection methods for use with students in my dissertation study.

Participation involves attending five 1.25-hour interviews. You would be compensated with $15 per session, paid in cash at the conclusion of each session; if you complete all five sessions you would receive an additional $25 bonus for completing the study, for a total compensation of $100.

If you're interested, you can send me an email at brow3821@umn.edu or you can also respond online at z.umn.edu/sampling. Please respond by **Friday, February 9**. If too many students sign up, not everyone will be able to participate and we will select students who represent a range of academic performance. You will be notified as to whether or not you have been selected to participate in the study.

Thanks and hope to hear from you!
Ethan Brown

IRB study number: 00002197

Appendix A2: Pilot recruitment reminder email

This is Ethan Brown, and I wanted to give you one last chance to participate in our pilot study of new ways to teach statistics. We know we can improve the way we teach, and you can help by trying out some new activities.

Participants would attend five 1.25-hour interviews, with compensation of $15 per session and an extra $25 for completing all five sessions. Let me know if you're interested at brow3821@umn.edu or z.umn.edu/sampling. If too many students sign up, not everyone will be able to participate and we will select students who represent a range of academic performance. You will be notified as to whether or not you have been selected to participate in the study.

Thanks!
Ethan Brown

IRB study number: 00002197

Appendix A3: Study recruitment first contact email

I'm Ethan Brown, a PhD candidate in Educational Psychology (advised by Associate Professor Robert delMas), and I'm researching new approaches to teaching statistics. I've developed some new activities similar to the ones you're currently using in EPSY 3264: Basic and Applied Statistics and was hoping you'd be willing to participate in a series of interviews to explore your thinking about these new activities. By participating, you'd help statistics teachers understand more about how students learn statistics and how we might improve the way we teach statistics.

Participation involves attending five 1.25-hour interviews. You would be compensated with $15 per session, paid in cash at the conclusion of each session; if you complete all five sessions you would receive an additional $25 bonus for completing the study, for a total compensation of $100.

If you're interested, you can send me an email at brow3821@umn.edu or you can also respond online at z.umn.edu/sampling. Please respond by **Friday, March 9.** If too many students sign up, not everyone will be able to participate and we will select students who represent a range of academic performance. You will be notified as to whether or not you have been selected to participate in the study.

Thanks and hope to hear from you!
Ethan Brown

IRB study number: 00002197

Appendix A4: Study recruitment reminder email

This is Ethan Brown, and I wanted to give you one last chance to participate in our study of new ways to teach statistics. We know we can improve the way we teach, and you can help by trying out some new activities.

Participants would attend five 1.25-hour interviews, with compensation of $15 per session and an extra $25 for completing all five sessions. Let me know if you're interested at brow3821@umn.edu or z.umn.edu/sampling. If too many students sign up, not everyone will be able to participate and we will select students who represent a range of academic performance. You will be notified as to whether or not you have been selected to participate in the study.

Thanks!
Ethan Brown

IRB study number: 00002197

Appendix A5: Consent form given to all participants

Title of Research Study: Growing Sampling Variability

Researcher: Ethan C. Brown, PhD Candidate in Educational Psychology; advised by Robert delMas, Associate Professor of Educational Psychology.

Supported By: This research is supported by the University of Minnesota.

Why am I being asked to take part in this research study?

We are asking you to take part in this research study because you have enrolled in EPSY 3264: Basic and Applied Statistics.

What should I know about a research study?

- Someone will explain this research study to you.
- Whether or not you take part is up to you.
- You can choose not to take part.
- You can agree to take part and later change your mind.
- Your decision will not be held against you.
- You can ask all the questions you want before you decide.

Who can I talk to?

For questions about research appointments, the research study, research results, or other concerns, call the study team at:

Researcher Name: Ethan C. Brown	Advisor name: Robert delMas
Researcher Affiliation: University of Minnesota	University of Minnesota
Phone Number: 612-625-2756	612-625-2076
Email Address: brow3821@umn.edu	delma001@umn.edu

This research has been reviewed and approved by an Institutional Review Board (IRB) within the Human Research Protections Program (HRPP). To share feedback privately with the HRPP about your research experience, call the Research Participants' Advocate Line at (612) 625-1650 or go to https://research.umn.edu/units/hrpp/research-participants/questions-concerns. You are encouraged to contact the HRPP if:

- Your questions, concerns, or complaints are not being answered by the research team.
- You cannot reach the research team.
- You want to talk to someone besides the research team.
- You have questions about your rights as a research participant.
- You want to get information or provide input about this research.

Why is this research being done?

We are seeking ways of improving student's understanding of the role of sample size in statistics. Students often have difficulty with understanding the role of sample size, so we developed a new approach based on new research in cognitive science and conceptual change. This research will help us understand and possibly improve the teaching of the role of sample size, and more generally will help researchers and teachers understand how different kinds of explanations may support students' understanding.

How long will the research last?

We expect that you will be in this research study for up to five weeks, or until you complete all five interview sessions.

How many people will be studied?

We expect between 2 and 50 people will be in this research study.

What happens if I say "Yes, I want to be in this research"?

Your participation would involve five 1.25-hour sessions with the researcher, Ethan C. Brown, in a room on the University of Minnesota—Twin Cities campus. These would be scheduled based on your and the researcher's availability.

These sessions would involve video-recorded one-on-one interviewing and tutoring with the researcher. Video-recording is required in order to participate in the research. The first session would consist of an interview to understand your current understanding of the role of sample size. The second through fourth sessions would consist of you working through a series of activities regarding the role of sample size, structured in a similar way to activities in EPSY 3264. The researcher would ask you questions and provide prompts to help you understand the material in these activities. The fifth session would consist of an interview to understand your understanding of the role of sample size after participating in the series of activities.

What happens if I do not want to be in this research?

You can decide to not participate in the research study and it will not be held against you.

What happens if I say "Yes", but I change my mind later?

You can leave the research study at any time and it will not be held against you. You may either choose to withdraw completely from the study and request that we destroy all data collected on you, or you may choose to discontinue future participation but still allow us to analyze and use the data we have collected so far.

Is there any way being in this study could be bad for me?

This study involves new approaches to teaching statistics. Because the approaches are new and being investigated for the first time, it is possible they could cause you confusion in EPSY 3264 and other statistical situations in your life. However, the approaches are based on educational theory and best practices in education and are expected to facilitate your understanding of the role of sample size in statistics.

Will being in this study help me in any way?

We cannot promise any benefits to you or others from your taking part in this research. However, possible benefits include a better understanding of the role of sample size in statistics, which is relevant both in EPSY 3264 and in many other statistical situations. Your participation in the study will also help us understand how to teach this difficult topic better, which may improve the instruction of EPSY 3264 and other statistics courses in the future.

What happens to the information collected for the research?

Efforts will be made to limit the use and disclosure of your personal information to people who have a need to review this information. We cannot promise complete secrecy.

Organizations that may inspect and copy your information include the IRB and other representatives of this institution.

Will I have a chance to provide feedback after the study is over?

The Human Research Protection Program may ask you to complete a survey that asks about your experience as a research participant. You do not have to complete the survey if you do not want to. If you do choose to complete the survey, your responses will be anonymous.

If you are not asked to complete a survey, but you would like to share feedback, please contact the study team or the Human Research Protection Program (HRPP). See the “Who can I talk to?” section of this form for study team and HRPP contact information.

Will I be compensated for my participation?

If you agree to take part in this research study, you will receive $15 for each session you attend for your time and effort. If you participate in all six sessions, you will receive an additional $25 bonus, for a total of $100 for all six sessions.

Payment you receive as compensation for participation in research is considered taxable income. If payment to an individual exceeds $600 in any one calendar year, the University of Minnesota is required to report this information to the Internal Revenue Service (IRS). Research payments to study participants exceeding $600.00 during any calendar year will result in a FORM 1099 (Miscellaneous Income) being issued to you and a copy sent to the IRS.

Statement of Consent for video-recording participation in Growing Sampling Variability. *Can we video-record you in this study?*

Please check one of the boxes and sign your name below.

□ Yes. The research team can record video of me during this study. Also, the researchers may use video clips from these video recordings when giving presentations of this research. I give my permission for the researchers to retain the video indefinitely and to use excerpts from my video in these presentations. Your video will be kept securely and will only be presented in the form of brief excerpts during professional presentations and lectures. The recording will be transcribed and the de-identified transcript may be published in entirety.

□ Yes. The research team can record video of me during this study. However, the researchers may NOT show the video or images from the video in any presentations. I do NOT give my permission for the researchers to retain the video of my participation. The video will be deleted at the conclusion of the study. The recording will be transcribed and the de-identified transcript may be published in entirety.

□ No. The research team may not record video of me. I do NOT give my permission to be video-recorded. This means I do not enroll in the study.

Question 1: Do you want the results of the study? □ Yes □ No

If yes, where should we send the information?

Address: ____

Email address: ____

Question 2: Do you want a copy of this consent form? □ Yes □ No

Your signature documents your permission to take part in this research.

____

Signature of participant Date

____

Printed name of participant

____

Signature of person obtaining consent Date

____

Printed name of person obtaining consent

Appendix B: Interview Protocols

Appendix B1: Pre-Interview

[Participant (P) and researcher (R) meet in lab testing room on University of Minnesota—Twin Cities campus. R gives P consent form; after the participant has finished reading silently, the R guides P through the consent form paragraph by paragraph. R checks with P whether they have any questions. If P consents to be in the study and to be videorecorded, the interview proceeds as below.]

[After P does so or refuses.] OK. I'm going to begin recording now. *[R starts recording equipment and ensures that it is working properly.]* Today I'll be asking you to solve several problems. I'm interested in the way that you reason about the problems, so I may ask questions to clarify your reasoning. If I ask a question, it doesn't necessarily mean that you're doing it wrong, or doing it right. I just want to make sure that I understand what you're saying.

To help us with this process, I would like to ask you to think aloud when you answer each question—to think and talk aloud as much as possible. At times I will remind you to think aloud as you answer a question and I might even ask you to explain something that you say so we can learn what you are thinking about for each question. Our purpose here is to learn about your understanding of the question and to understand your reasoning for your responses. We're most interested in how you came to an answer, and what made you choose an answer, rather than what the answer itself is.

Please remember that I do want to hear all of your opinions and reactions. Do not hesitate to let us know that something is unclear, difficult to answer, or does not apply to you. Do you have any questions before we start?

[Wait for questions.]

If you have any questions or concerns at any time in the interview, feel free to stop me and ask.

Practice questions

Let's start with a couple of practice questions. These are not statistics problems, just questions to practice the process of thinking aloud.

[Hand Practice Question 1 to P.]

This page contains a problem. Please read the problem aloud to me and then talk to me about your interpretation of the problem and your thinking about how you would respond.

Practice Question 1. How many windows are there in the place where you live?

Potential Follow-up Questions:

Do you have two layers of windows in some places where you live? Are you counting both layers as a single window or as two windows? Why?

Do some of your windows have multiple parts? Do you count each pane separately, or all as one window? Why?

How confident are you of your answer?

[Takes paper away.]

OK. Can you repeat back to me the question I just asked you?

Great. As you can see, there's different ways to count the number of windows. When I ask you these questions, I'm trying to understand how you're making sense of the question. It's OK for your thinking to change, or to stay the same as we discuss the question… it's *your* thinking. My questions are just like the questions I'll soon be asking you about statistical situations. Let's try another practice question.

Practice Question 2. How difficult was it for you to get here to do the interview today?

Very difficult, somewhat difficult, a little difficult, not at all difficult

1. *Follow-up questions:*
2. *What was the easiest part of getting here today? Why was it easy?*
3. *What was the hardest part of getting here today? Why was it hard?*
4. *How confident are you of your answer?*

Great. That's exactly the kind of explanations we're looking for.

[Takes paper away.]

Now, can you repeat back to me the question I just asked you?

I'll be asking some statistical questions now, but I want you to think of these questions in the same way as the windows and difficulty questions. I'm not your statistics teacher who's hoping for a correct answer. I'm just a researcher who wants to understand how you're making sense of these problems.

Again, read the question aloud to me and describe aloud what you're thinking as you're answering the question. I may remind you to keep talking to keep the conversation moving.

Pre-questions

R hands packet to P. The packet contains the Hospital, Referendum, Candies, Batting Averages, Post Office, and Casino problems, one on each page, so that the participant can refer to the questions and write anything if they so choose.

P reads question, thinks aloud about the solution. I probes for reasoning, including asking "why" questions.. Then, take question away and ask: "Now, repeat back to me the question I gave to you." Give the next question to P and repeat.

Post Office Simulation

People often have difficulty understanding problems like this, so I'd like to provide some visuals to help you understand exactly what the question is asking.

[Open up heights_model.tp. The file a data table of heights and weights. Save As ***S*#_*01-01_heights_model.tp filling in the subject number.]*

Here is a dataset of individual heights and weights similar to those we described in the problem. Imagine that these are the heights of 18-year-old men in the city that has Post Office A and Post Office B. Now, I'd like you to plot these heights to get a sense of how they're distributed.

[P plots heights. Assist as needed. The distribution is very bell-shaped with a peak at 68.]

Great. Now, can you tell me what the overall average height is?

The average here is the same as what is listed in the problem – 5 feet 8 inches is 68 inches. What proportion of men in the town have a height greater than 68 inches?

[Probe, tech assist as necessary.]

How about 72?

Now, imagine that we will pick 10 people randomly from the city. How could you use TinkerPlots to randomly draw 10 people from this group?

This file has all the 18-year-old men in the city in a TinkerPlots mixer. Let's look at one sample of 10 people.

[Coach P to draw one sample, plot and show the mean.]

Is the mean going to be the same every time we draw a sample?

The problem says that Post Office A, 10 men registered every day for a year, and every day they noted the average height of the 10 men. Set up a TinkerPlots model for taking the average of 10 men and recording the result.

[Coach towards collecting statistic on the existing setup. Don't use the "collect" button yet: we set up plotting first in order to make sure P understands what each plot represents.]

Let's take the average another time and create a plot of the averages.

[Press Run button again to get another average so that TP can create plot.]

Now, we have three plots here. What does this dot in this plot represent? *[Click on a dot in the original plot of all 18-year-olds.]*

How about this dot? *[Click on a dot in the plot of a sample of all 18-year olds.]*

Finally, how about this dot? *[Click on a dot in the plot of averages.]*

[Provide correct answers if incorrect.]

I've guided you through the TinkerPlots setup here, but now I'd like to return to my role as a researcher who's interested in how you're making sense of this situation.

Imagine that we repeat this process every day for a year, just like in the problem. What proportion of the sample means—the dots in this graph of averages—would you expect to be over 72 inches?

Sketch a graph of what you think the distribution of sample means might look like.

Great, now let's look at what actually happens.

[Direct P to turn off animation and minimize windows for speed. If P is concerned about business days, say that there are about 260 business days in a year; if they use 365 for every single day, do not mention this complexity.]

Is this surprising, or what you expected? Sketch a graph of what this distribution looks like.

Let's think about Post Office B, now – where 100 people registered each day. Imagine that we generate this same plot of average, now for averages of 100 instead of 10. What proportion of the sample means—the dots in this graph of averages—would you expect to be over 72 inches? *[Probe for why]*

Sketch a graph of what you think the distribution of sample means might look like. *[Probe for why.]*

Great, now again let's look at what actually happens.

[Guide P to delete current history collection.]

Sketch the actual graph, now.

What do you notice?

Why are the two distributions different?

Appendix B2: Growing Sample Proportions and Means

Growing Sample Proportions: Physical Simulation

Here's a box with one orange block and one blue block. We're going to shake the box, take out a block, record it on this sheet, and then put the block back into the box. We're also going to mark how many blue we've seen so far, and the proportion blue we see so far.

[Shake box thoroughly, draw one: say e.g. "This one was blue, so you would write 'B' here. That also means we have one blue so far. All of the blocks so far are blue, so the proportion is 1.00 or 100%." Repeat. Check to make sure P understands.]

You can use this calculator if you need to calculate anything.

Before we start, what do you expect this table to look like? *[Probe for each column in table.]*

OK. Let's draw the 10 blocks.

[P shakes box and writes down on table in sheet.]

What do you notice about the proportion of blue blocks as we draw more blocks?

What do you think would happen if we kept on drawing more blocks?

If we started over and did this again, what would you expect to be similar? What would you expect to be different?

Growing Sample Proportions: TinkerPlots™

We're going to be exploring a new kind of plot, a *sample size plot*. Let's do an example in TinkerPlots.

[Open 02-01_Proportion_vs_Sample_Size.tp]

This table shows a sample size, and the proportion at that sample size, similar to the table you created. These numbers were generated randomly just for demonstration. Let's create a plot of this data.

[Guide P to plot proportion on x-axis, sample size on y-axis, borderless icons, line.]

So, the sample size plot shows how the proportion changes as sample size increases.

What do you think the sample size plot would look like for the proportion blue here?

Please sketch what you think the plot might look like growing from n = 1 to n = 50.

I'd like for you to describe for me: How could we set up the block situation in TinkerPlots?

Let's see what that looks like.

[Guide P to use Repeat +1 to incrementally add on to the sample and to erase old ones]

OK – let's plot this and look at the proportion blue as the sample size increases. Let's go up to 50 to see what happens.

[P goes to 50, then stops.]

What do you notice? Can you sketch here the plot of what you saw happening with the proportion by sample size?

Growing Sample Means: 0-1

We're going to change one thing. We're going to model the same situation, shaking the box and drawing one block. But now, every time we draw blue, we score it as 1, and every time we draw orange, we score it as 0. *[Shake the box, draw one.]* So, we would track this as a *[0 or 1, depending on whether orange or blue]*. How could you set this up in TinkerPlots?

OK. Now, we're going to be interested in the *mean*, instead of the proportion orange, as the sample size increases. Can you sketch what you think the sample size plot would look like for the mean?

We'll add some formulas so we can actually create the plot in TinkerPlots. *[Coach P to add column with total number, using formula prev(Total) + attribute, and Sample Size, and Mean.]*

What do you notice? Sketch the plot that you actually saw.

What would happen if we started over and did this again?

Growing Sample Means: 0-1-1

Now, suppose there were one orange and TWO blue blocks in the box. Again, each blue block is scored 1 point and each orange block is scored as 0 points. *[Bring out box and demonstrate.]* What would you expect to see in terms of the mean score as you keep drawing blocks? What do you expect would be different from before? What would be similar?

Please draw what you'd expect the sample size plot to look like.

How can you change the model in TinkerPlots?

Let's see what that looks like.

Sketch the sample size plot that you actually saw.

What do you notice looking at this plot? What is similar to when you had one orange and one blue block? What is different?

What would happen if we started over and did this again?

Growing Sample Means: Cat Factory 1

Do you remember an activity from class where you created a cat factory?

We're going to create another cat factory now, where you can create cats of different lengths.

[Open Template_02-03_cat_01.tp]

Here, you can see that the cats have different possible lengths between 6 and 32. The bars represent how likely each length of cat will be, and so that all cat lengths are equally likely. How reasonable does this seem as a model for creating cats?

[If not reasonable:] OK. Draw in the Sampler window whatever likelihood seems reasonable.

Click the Run button a couple times. This file is set up so that the mean appears here, right next to the cat length. You can see a plot of the cat length in the sample here, but you can also see here a plot of the means by the sample size. So, this square shows that the mean at sample size 2 was ______.

What do you think the graph would look like if you kept growing the sample? Why do you think this happens? Please sketch your sampler and what you'd expect to see for the sample size graph.

[Hand next sheet for sketching graphs.]

What did you notice? What would happen if we started over and did this again?

Growing Sample Means: Cat Factory 2

What could you change in the sampler to make the graphs look different? *[Coach towards changing the relative heights of the bars.]* Why might that make the graphs look different?

Again, sketch your sampler and what you'd expect to see for the sample size graph.

[Hand next sheet.]

OK, let's see what actually happens.

What did you notice? What do you notice is similar to your previous graph? What is different?

[Probe about "spikes" if this hasn't come up yet]

Examples of the types questions and follow-up questions that students will be asked during the task-based interview:

1. *What is happening as the sample size increases? Why?*
2. *How much does the mean change when you add on a new sample?*
3. *How does adding an additional sample value change the sample mean? Does it have the same amount of influence on a large sample as it does on a small sample? Why or why not?*
4. *What is similar between different times you grow the sample? What is different? Why do you think this is true?*
5. *How much is the mean moving around at low sample sizes vs. high sample sizes?*
6. *What value does the mean appear to approach as the sample size increases? Why?*
7. *Notice there are occasional spikes in the sample size graph. What do these spikes correspond to? What changes about the spikes as sample size increases? Why?*

Appendix B3: The Mystery Mean & Growing More Means

The third interview did not have an interview protocol. See sections 3.3.6, p. 100 and 3.3.7, p. 103 for details about the tasks and structure of this interview, and Appendix D3: The Mystery Mean & Growing More Means, p. 438 for the full transcript with participant S.

Appendix B4: Growing Possibilities & Many Means

Growing Possibilities: 0-1 Building Blocks

N=1 We're going to think again about the situation where we're drawing blocks from the box. Here I have one white block and one black block in the box. Just like before, I shake the box, randomly pick a block, and put that block back in the box. Here, the black blocks is scored as 1, and the white blocks is scored as 0. Again, we're interested in the *average score* at different sample sizes.

Now, we're going to create some graphs of the different possible outcomes and the means of those outcomes. Here at n = 1, what are the possible outcomes?

What are the means of each of those possible outcomes?

So, we'll place the white block at 0 and the black block at 1, like this. *[Demonstrates]*

N =2 Now, we want to look at the all the possible outcomes when we draw two blocks from the box. We can represent the different outcomes by sticking the blocks together. So, if we draw a black block and then a white block, we can stick them together like this *[demonstrates]*. What would the mean be if we draw one black block—a 1—and then a white block—a 0?

So, we'll place this block there.

Now, put together blocks to represent drawing two white blocks, and position those on the paper in the same way.

Are these all the possibilities that end in a white block?

How do these possibilities ending with a white block relate to the possible outcomes at n = 1? *[probe as needed to explore connection between the two levels.]*

Now, please put together all the possibilities that end in a *black* block and position them on the paper.

How do these possibilities ending with a black block relate to the possible outcomes at n = 1? *[probe as needed to explore connection between the two levels.]*

Let's stack all the possibilities together neatly, just like we do in TinkerPlots. What do you notice? Which means do you think are the most likely? Which means are least likely?

N = 3. Let's do the same thing for n = 3. Create all the outcomes that end in a white block and position them on the paper.

How do these possibilities ending with a white block relate to the possible outcomes at n = 2? *[probe as needed to explore connection between the two levels.]*

Now, please put together all the possibilities that end in a *black* block and position them on the paper.

How do these possibilities ending with a black block relate to the possible outcomes at n = 2? *[probe as needed to explore connection between the two levels.]*

What do you notice? Which means do you think are the most likely? Which means are least likely?

N = 4. Finally, let's look at n = 4. Create all the outcomes that end in a white block and position them on the paper.

How do these possibilities ending with a white block relate to the possible outcomes at n = 3? *[probe as needed to explore connection between the two levels.]*

Now, please put together all the possibilities that end in a *black* block and position them on the paper.

How do these possibilities ending with a black block relate to the possible outcomes at n = 3? *[probe as needed to explore connection between the two levels.]*

What do you notice? Which means do you think are the most likely? Which means are least likely?

What would you expect to happen if we kept increasing the sample size? At n = 10, would 0 be more or less likely than here at n = 4? What percentage of the outcomes would you expect to be near the center at n = 10? At n = 50?

Growing Possibilities: 0-1 Sample Size Plots

I'm going to pick one of the samples here in this plot and I'd like you to draw the sample size plot of the sample. These are the graphs we made earlier, where the mean is on the x-axis and the sample size increases as you go up the y-axis. What would the sample size plot look like for this block between n=1 and n=4 for this block? (BBBB) Just draw the first part between 1 and 4 on this graph. Back with the other blocks, draw a line between this block and where this sample would be at n =3, n=2, n=1. Now, draw what do you think might happen as the sample size grows to n = 25.

What would the sample size plot look like for n = 1 to n = 4 for this block? (WBWW) Back with the other blocks, draw a line between this block and where this sample would be at n =3, n=2, n=1. Draw what do you think might happen as the sample size grows to n = 25.

What would the sample size plot look like for n = 1 to n = 4 for this block? (BBWB) Back with the other blocks, draw a line between this block and where this sample would be at n =3, n=2, n=1. Draw what do you think might happen as the sample size grows to n = 25.

What would the sample size plot look like for n = 1 to n = 4 for this block? (WBBB) Back with the other blocks, draw a line between this block and where this sample would be at n =3, n=2, n=1. Draw what do you think might happen as the sample size grows to n = 25.

What do you see happening to the distribution as the sample size increases?

Growing Possibilities: 0-1-1 Building Blocks

N=1 Now, we're going to add a red block to the box, but the red block is also scored one. Just like before, I shake the box, randomly pick a block, and put that block back in the box. Now the red block and the black block are scored as 1, and the white blocks is scored as 0.

Here at n = 1, what are the possible outcomes?

What are the means of each of those possible outcomes?

So, we'll place the white block at 0 and the red and black block at 1, like this. *[Demonstrates]*

N =2 Now, we want to look at the all the possible outcomes when we draw two blocks from the box. We can represent the different outcomes by sticking the blocks together. So, if we draw a black block and then a white block, we can stick them together like this *[demonstrates]*. What would the mean be if we draw one black block—a 1—and then a white block—a 0?

So, we'll place this block there.

Now, put together blocks to represent drawing two white blocks, and position those on the paper in the same way.

Are these all the possibilities that end in a white block?

How do these possibilities ending with a white block relate to the possible outcomes at n = 1? *[probe as needed to explore connection between the two levels.]*

Now, please put together all the possibilities that end in a *black* block and position them on the paper.

How do these possibilities ending with a black block relate to the possible outcomes at n = 1? *[probe as needed to explore connection between the two levels.]*

Now, please put together all the possibilities that end in a *red* block and position them on the paper.

How do these possibilities ending with a red block relate to the possible outcomes at n = 1? *[probe as needed to explore connection between the two levels.]*

Let's stack all the possibilities together neatly, just like we do in TinkerPlots. What do you notice? Which means do you think are the most likely? Which means are least likely?

N = 3. Let's do the same thing for n = 3. Create all the outcomes that end in a white block and position them on the paper.

How do these possibilities ending with a white block relate to the possible outcomes at n = 2? *[probe as needed to explore connection between the two levels.]*

Now, please put together all the possibilities that end in a *black* block and position them on the paper.

How do these possibilities ending with a black block relate to the possible outcomes at n = 2? *[probe as needed to explore connection between the two levels.]*

Now, please put together all the possibilities that end in a *red* block and position them on the paper.

How do these possibilities ending with a red block relate to the possible outcomes at n = 1? *[probe as needed to explore connection between the two levels.]*

What do you notice? Which means do you think are the most likely? Which means are least likely?

N = 4. Finally, let's look at n = 4. Create all the outcomes that end in a white block and position them on the paper.

How do these possibilities ending with a white block relate to the possible outcomes at n = 3? *[probe as needed to explore connection between the two levels.]*

Now, please put together all the possibilities that end in a *black* block and position them on the paper.

How do these possibilities ending with a black block relate to the possible outcomes at n = 3? *[probe as needed to explore connection between the two levels.]*

What do you notice? Which means do you think are the most likely? Which means are least likely?

Now, please put together all the possibilities that end in a *red* block and position them on the paper.

How do these possibilities ending with a red block relate to the possible outcomes at n = 1? *[probe as needed to explore connection between the two levels.]*

What would you expect to happen if we kept increasing the sample size? At n = 10, would 0 be more or less likely than here at n = 4? What percentage of the outcomes would you expect to be near the center at n = 10? At n = 50?

Growing Possibilities: 0-1-1 Sample Size Plots

What would the sample size plot look like for this block between n=1 and n=4 for this block? (WRW) Just draw the first part between 1 and 4 on this graph. Back with the other blocks, draw a line between this block and where this sample would be at n =3, n=2, n=1. Now, draw what do you think might happen as the sample size grows to n = 25.

What would the sample size plot look like for n = 1 to n = 4 for this block? (RRB) Back with the other blocks, draw a line between this block and where this sample would be at n =3, n=2, n=1. Draw what do you think might happen as the sample size grows to n = 25.

What would the sample size plot look like for n = 1 to n = 4 for this block? (BWR) Back with the other blocks, draw a line between this block and where this sample would be at n =3, n=2, n=1. Draw what do you think might happen as the sample size grows to n = 25.

What would the sample size plot look like for n = 1 to n = 4 for this block? (BBW) Back with the other blocks, draw a line between this block and where this sample would be at n =3, n=2, n=1. Draw what do you think might happen as the sample size grows to n = 25.

What do you see happening to the distribution as the sample size increases?

Growing Many Means: 0-1

Here's a file to explore both the distribution off means and the sample size plots. Again, I can draw 0 or 1, just like we saw with the blocks. Now, the slider here determines the sample size that we'll see. After I draw a sample, you can see that the sample size is 1 and this window is plotting the mean at that sample size. If I bring it up to n=2, now I'm showing the means of sample size 2. If I click here you can see this is a time where I drew a 1 and then a 0. Does that make sense?

Ok. I'd like you to change the slider to 4.

There's more we can show. (Show hidden) This plot shows the individual sample values, e.g… and here we see the mean by sample size plot for the sample corresponding to the number that appears in the slider, which is also highlighted in black up here. To figure out the sample number, we can click on a sample in the top plot and hat sample gets highlighted in the table. Then we can just move the slider to or type in that number to show it.

Let's do an example – click on a sample mean and then show the plots related to that sample.

Great. Let'slook back at the sample size plots you drew. Can you find a sample that matches the beginning of your sample size plot?

Ok. Let'ssee what happens as it grows to n=25. Press the play button for the n slider. Stop. Was that what you expected? What happened to the plot of means?

Go back to n=4 and find a sample that matches the beginning of the next sample size plot. Let's see what happens as this sample grows to n = 25.

Again, let's go back to n=4 and find the matching sample. Let's grow to n=25.

Again, let's go back to n=4 and find the matching sample. Let's grow to n=25.

What do you think will happen if we grow to n = 100? Describe what happened. Was that what you expected? Why?

Let's go to n = 15 now. What do you think will happen to these samples on the extremes? Ok, I'd like for you to select all of those. *[Coach to use Shift-select to select both sides.]* Let's look at what happens only to those. *[Coach to "Hide Unselected Cases."]* Let's also follow one of those samples as it grows.

Describe what happened. Was that what you expected? Why?

Growing Many Means: 0-1-1

OK, now we're going to add another 1 into the sampler. What do you think will change when we do that? *[repeat process of finding sample size plots, growing to 25, selecting extremes, and growing to 100].*

Appendix B5: Post-Interview

Post-questions

Once again, we're going to be solving some problems today. Just like on the first day, I'd like you to read the problem aloud, and to think aloud when you answer each question—to think and talk aloud as much as possible. I may remind you to keep talking to keep the conversation moving.

Any questions before we begin?

Pre-questions, revisited

Now, I'd like to return to the problems we did on the first day. Again, please read each problem aloud and think aloud as you solve the problem.

Let's look back at how you originally answered these problems. I'll put each in front of you. *[Place original and new response to pre-question side-by side]* Let me know what has changed, and what has stayed the same, in how you reason about the problem. *[Repeat for all pre-questions.]*

Comparing representations

As you know, we've looked both at some sample size plots and looked at combinations, at the distributions of means. *[pull up participant's TinkerPlots™ files]* Do you see connections between the sample size plot and the Geology problem? How about the combination plot *[show 0-1-1 with combos]*? *[Repeat for Coin Flip problem]*

Interview closing

What did you think these activities/tasks were getting at? Did you notice anything about your thinking change?

How would you explain to someone the impact of sample size on sampling variability?

Any comments about your participation in the study? What did you like about the activities? What didn't you like as much about the activities? Do you havbe any questions for me?

Appendix C: Pre- and post-interview questions

Pre- and post-interview questions administered on paper are reproduced below. TinkerPlots™ files and blank graphs will be made available through the Data Repository at the University of Minnesota (DRUM).

Appendix C1: Pre-Interview

Each of the pre-inverview questions were originally presented on separate pieces of paper.

Practice Question 1. How many windows are there in the place where you live?

Practice Question 2. How difficult was it for you to get here to do the interview today?

Circle one: Very difficult, somewhat difficult, a little difficult, not at all difficult

Hospital. A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50% of all babies are boys. The exact percentage of baby boys, however, varies from day to day. Sometimes it may be higher than 50%, sometimes lower.

For a period of 1 year, each hospital recorded the days on which more than 60% of the babies born were boys. Which hospital do you think recorded more such days? Explain your reasoning,

A. The larger hospital

B. The smaller hospital

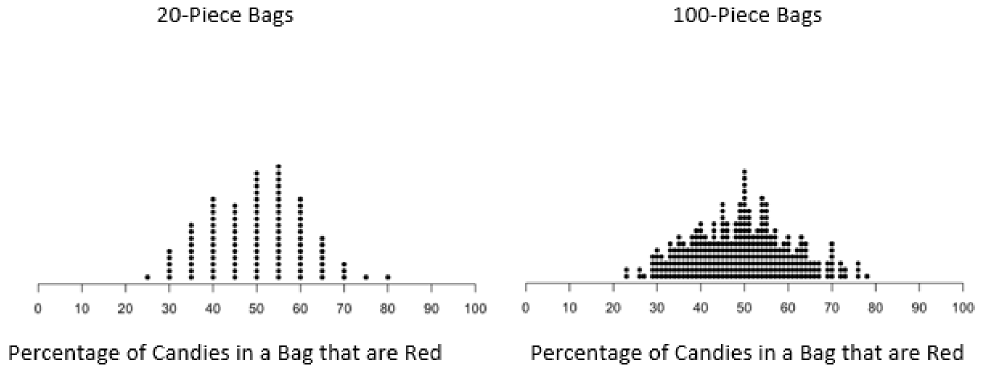
C. About the same

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

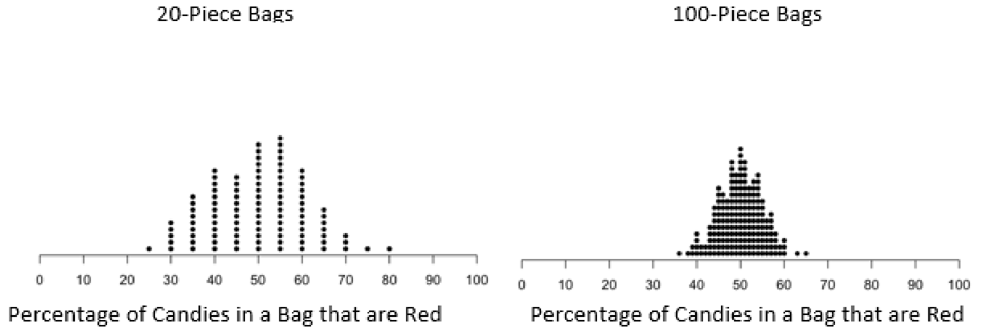
Candy. Imagine a candy company that manufactures a particular type of candy where 50% of the candies are red. The manufacturing process guarantees that candy pieces are randomly placed into bags. The candy company produces bags with 20 pieces of candy and bags with 100 pieces of candy.

Which pair of distributions (below) most accurately represents the variability in the percentage of red candies in an individual bag that would be expected from many different bags of candy for the two different bag sizes?

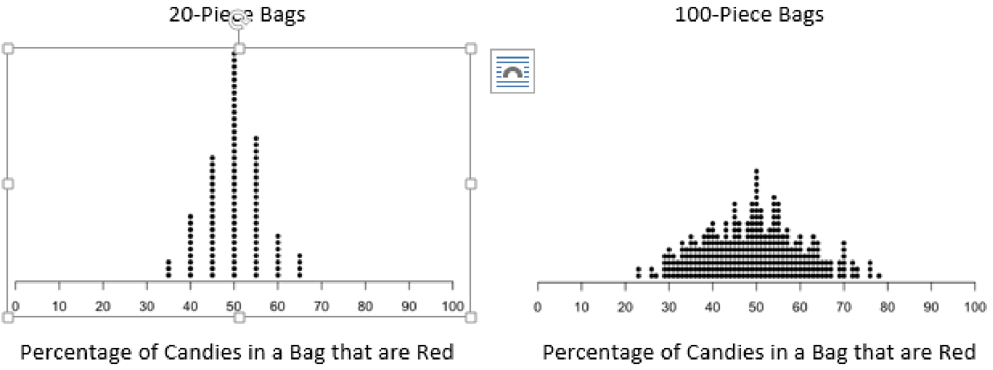
a.



b.



c.



Batting average. In baseball, players are often evaluated by their "batting average", which is the proportion of times that they hit the ball. In 2016, the batting average for the entire league was .255. After the first few baseball games in the season, several players may have a batting average of .450. However, those players will usually have a batting average that is lower than .450 by the end of the season. Why could this be true? Explain.

Post office. When they turn 18, American males must register for the draft at a local post office. In addition to other information, the height of each male is obtained. The national average height of 18-year-old males is 5 feet, 8 inches.

Every day for one year, 10 men registered at post office A and 100 men registered at post office B. At the end of each day, a clerk at each post office computed and recorded the average height of the men who had registered there that day.

Which would you expect to be true? (circle one)

1. The number of days on which the average height was 6 feet or more was greater for post office A than for post office B.

2. The number of days on which the average height was 6 feet or more was greater for post office B than for post office A.

3. There is no reason to expect that the number of days on which the average height was 6 feet or more was greater for one post office than for the other.

Casino. You work for the state casino regulation committee. Your job is to ensure that casinos are accurately reporting to customers the average winnings from slot machines. Suppose one slot machine pays out \$0, \$1, or \$20 on each game, and the machine claims that the average payout is \$0.90. You can play the slot machine as many times as you want, but it costs money each time. Construct a proposed strategy for determining whether the slot machine's claim is accurate.

Appendix C2: Post-Interview

Each post-interview question was administered on a separate page. Then, all the pre-interview questions, above, were re-administered.

Geology. In a geology course, an instructor has her students weigh a metal disk several times on the same scale. The scale is not completely accurate and is slightly inconsistent from weighing to weighing. However, the scale is equally likely to read above the true weight as it is to read below the true weight.

The class is divided into two teams, led by Jaiden and Paulina. Jaiden's team decides to weigh the disk 20 times, then compute and record the average of the 20 weighings. Paulina's team decides to weigh the disk 5 times, then compute and record the average of the 5 weighings.

Suppose the true weight of the disk is 2 pounds. All the students are experienced with using the scale, and record the average weight that they found. Each student also notes whether their average was above 2.2 pounds.

Which of the following would you expect to be true about the students' average recorded weights?

a) More of Jaiden's team (20 weighings) will have average weights above 2.2 pounds.

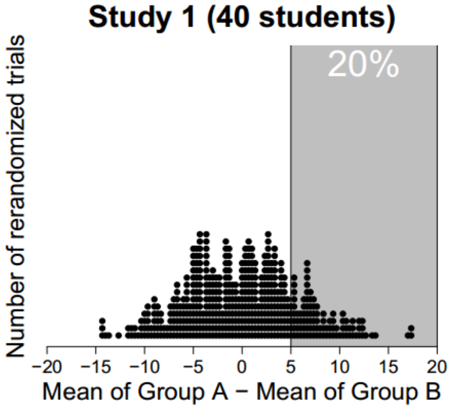
b) More of Paulina's team (5 weighings) will have average weights above 2.2 pounds.

c) There is no reason to think that either team's weighings will be more likely to have average weights above 2.2 pounds.

Factory. Bert has a job checking the quality of glass bottles made in a bottle factory that makes 90 bottles every day. Overall, the machine makes perfect bottles about 80% of the time. Bert has noticed that on some days, all of the first 10 bottles are perfect. However, Bert has also noticed that on such days, the overall percentage of perfect bottles is usually similar to days when some of the first 10 bottles are imperfect.

Why do you suppose the percentage of perfect bottles is usually not much better on days where the first 10 bottles are perfect?

Exam Preparation. Consider an experiment, Study 1, where a researcher wants to study the effects of two different exam preparation strategies on exam scores. Forty students volunteered to be in the study, and were randomly assigned to one of two different exam



preparation strategies, 20 students per strategy. After the preparation, all students were given the same exam. The researcher calculated the mean exam score for each group of students. The mean exam score for the students assigned to preparation strategy A was 5 points higher than the mean exam score for the students assigned to preparation strategy B. The researcher ran a randomization test for the difference in means and plotted the mean differences for 500 rerandomized trials:

The researcher noted that 20% of the rerandomized mean differences for Study 1 were greater than 5 points, and so the *p*-value for the mean difference in Study 1 was 0.20.

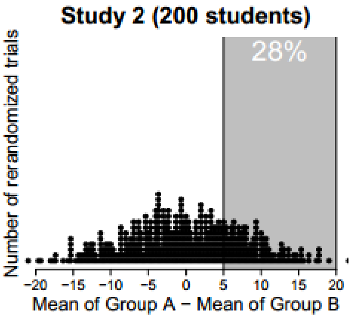
Imagine another study, Study 2, where 200 students participated, 100 in each group. In Study 2, the mean exam score for the students assigned to preparation strategy A was, again, 5 points higher than the mean exam score for the students assigned to preparation strategy B.

This table summarizes the important information about the two studies:Appendix C3:

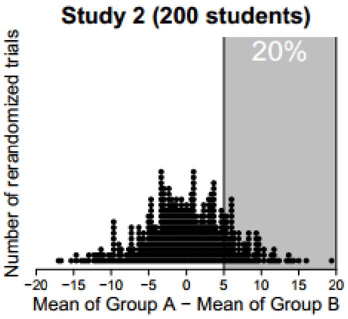
Study	**Sample Size**	**Mean difference**	***p*-value**
1	40	5 points	0.20
2	200	5 points	**?**

Which distribution of mean differences for 500 rerandomized trials (below) most accurately represents the expected *p*-value for the study where there were 200 students?

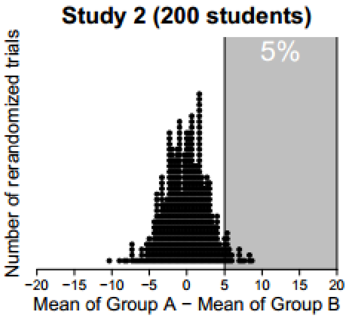
a)



b)



c)



Coin Flips. Two groups of students are flipping coins and recording whether or not the coin landed heads up. One group of students flips a coin 50 times and the other group of students flips a coin 100 times. Each student notes down the percentage of heads of all their flips. Which group will have more students who get more than 52% of their coin flips heads up? Explain.

Working Choices. An economist was interested in whether Americans would still work full-time even if they were provided with guaranteed unearned income from the government. She cited a recent study of 3,000 Americans, randomly sampled from the top 1% wealthiest Americans. Although everyone in the sample could afford to live comfortably without working, about 92% still worked full-time jobs. Therefore, she concluded, most Americans will still work full-time jobs even if the government provided a guaranteed unearned income.

Comment on the economist's reasoning. Is it basically sound? Does it have weaknesses?

Appendix D: Complete Transcript for Participant S

Gray numbers to the left of each line indicate line numbers in the original transcript. Times in the original video are indicated in brackets approximately every 10 lines of speech. Blank lines are kept in for consistency since line numbers in text and analysis are based on these transcripts. See beginning of Chapter 1 for transcription conventions. All interviews were transcribed from the screen recording, which had better audio, except for the beginning of the Post-Interview due to technical issues.

Appendix D1: Pre-Interview

1.1: Introduction

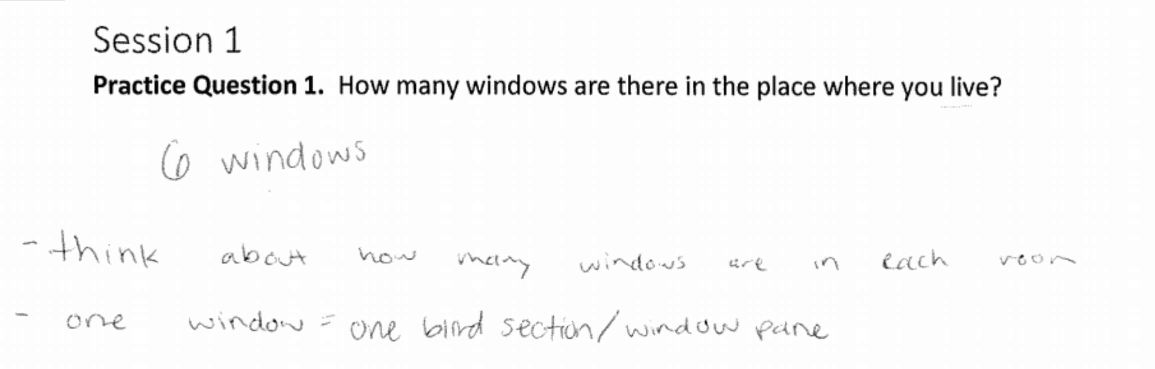
1.2: I:
1.3: [00:00:08] So, today I'll be asking you to solve
1.4: several problems. I'm interested in the
1.5: way that you reason about the problems.
1.6: And so I may ask questions to clarify
1.7: your reasoning. Um, if I ask a question, it
1.8: doesn't necessarily mean that you're
1.9: doing it wrong, or doing it right. I just
1.10: want to make sure that I understand what
1.11: you're saying.
1.12: Uh to help us with this process I'd like to
1.13: [00:00:32] ask you to think aloud when you answer
1.14: each question, to think and talk aloud as
1.15: much as possible. At times I'll remind
1.16: you to think aloud as you answer a
1.17: question, and I might even ask you to
1.18: explain something that you say so that
1.19: we can learn what you're thinking about
1.20: for each question. Our purpose here is to
1.21: learn about your understanding of the
1.22: question, and to understand your
1.23: [00:00:55] reasoning for your responses. We're most
1.24: interested in how you came to an answer,
1.25: and what made you choose an answer,
1.26: rather than the answer itself. So,
1.27: please remember I do want to hear all of
1.28: your opinions and reactions. S-- don't -- do not
1.29: hesitate to let us know if something is

unclear, difficult to answer, or doesn't apply to you. Do you have any questions before we start?

S: No.

Practice questions

I: Okay. If you have any [00:01:20] questions or concerns at any time in the interview, feel free to stop me and ask. So let's start with a couple practice questions. Um, these are not statistics problems /c/, just questions to practice the process of thinking aloud. So, um, this page contains a problem. Please read the problem aloud to me and then talk to me about your interpretation of the problem, and your thinking about how [00:01:46] you would respond.



Session 1

Practice Question 1. How many windows are there in the place where you live?

6 windows

- think about how many windows are in each room

- one window = one blind section/window pane

S: So the question is how many windows are there in the place where you live? So in my apartment, not my house, but -- I would go first, like, to each room, thinking about what -- what -- how many windows are in each room. So in my room, there's one. My roommates there's two, so

1.59: that's three, and then our main living
1.60: area, there's also one. And then in my
1.61: other roommates, there's one more window,
1.62: [00:02:10] and then the other room, there's another
1.63: one. So there's six windows.
1.64:
1.65: I: Okay, great.
1.66: And so what's -- um, what's counting as one
1.67: window for you?
1.68:
1.69: S: I'd probably say, like,
1.70: either, like, one single blind, or, um, in the
1.71: windows in the rooms, um, it like slides up
1.72: to open up the window, so whatever -- how
1.73: many sections there are. So.
1.74:
1.75: I: Okay. Okay. Gotcha.
1.76: So -- so tell me more -- tell me a little bit
1.77: more about the ones with sections?
1.78:
1.79: S: Yeah,
1.80: [00:02:43] so. Like, in my room it's just, like, one
1.81: windowpane and then, like, there's just
1.82: one part that goes up, like, there's no --
1.83: like there's not like a second window
1.84: next to it or anything. [OK] So.
1.85:
1.86: I: Gotcha, gotcha. Um, so -- um,
1.87: so part of what we're getting at, here, is
1.88: there's different ways you could count
1.89: the number of windows. [mh] Um, so when I ask
1.90: you these questions, I'm just trying to
1.91: understand how you're making sense of
1.92: [00:03:09] the question. Um, it's okay for your thinking
1.93: to change or to stay the same as we
1.94: discuss the question. It's your
1.95: thinking. Uh, my questions are just like the
1.96: questions I'll soon be asking you about
1.97: statistical questions -- about statistical
1.98: situations. Um, so could you write your
1.99: answer, first.

S: Should I just put, like, the number, or...

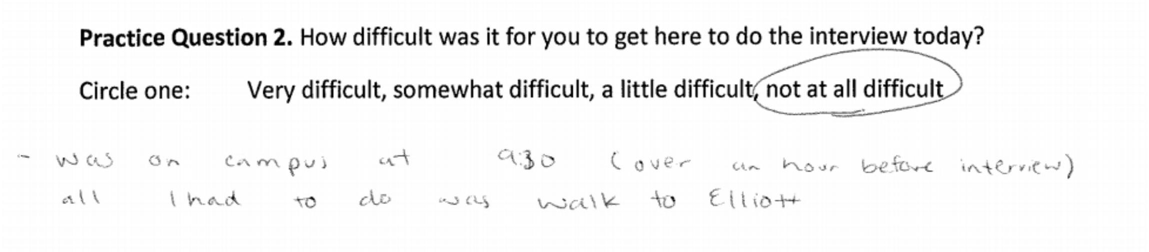
I: Yeah [OK], you can put the number. And then, um, briefly just a couple bullet [00:03:39] points, um, describing your reasoning for that answer. Okay, great. Um and how confident are you in your response to this question?

S: I'd say pretty confident /c/

I: OK, great. And finally, um can you repeat back to me the written problem?

S: So it was how many windows are there in your house -- or where you live [OK] so.

I: OK, great. So let's try another practice question.



Practice Question 2. How difficult was it for you to get here to do the interview today?

Circle one: Very difficult, somewhat difficult, a little difficult, [✓ not at all difficult]

- was on campus at 9:30 (over an hour before interview)
all I had to do was walk to Elliott

S:

[00:04:51] Okay. So the question is how difficult was it for you to get here to do the interview today? And then it says to circle one: very difficult, somewhat difficult, a little difficult, and not at all difficult. So I'm currently -- I moved for Spring Break, I'm back at my house. [mh] S-- but, I had a meeting with one of my professors this morning, so I just took the bus here which was really easy, I [00:05:13] just kind of sleep on the bus [mh] and then -- so I've been on campus since around like 9:30 or so. So, it was pretty easy, I just went back to my apartment then walked here, so it wasn't very difficult. /c/

I: Okay, and so which, uh, which answer would that lead you towards?

S: Probably not at all difficult.

I: Okay. And -- um, okay. And so, um, can you put another -- a couple bullet points explaining your [Yeah] reasoning. [00:06:14] Okay, great. So, I'll be asking -- oh and how confident are you in your answer to this?

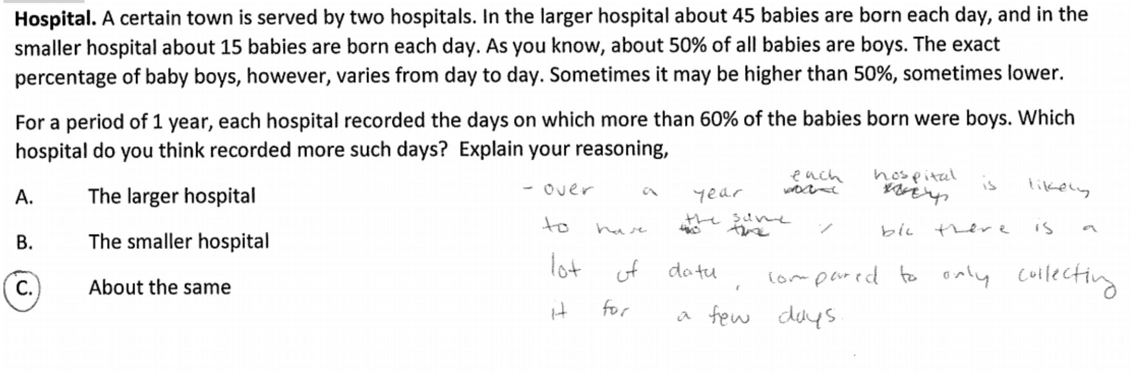
S: Very confident. /c/

I: Very confident. And um can you repeat back to me the written problem?

S: It was how difficult to get here today, /c/ for the interview?

I: Okay so I'll be asking some statistical questions, now. But I want you to think of these questions in the same way as the windows and difficulty [00:06:45] questions. I'm not your statistics teacher, who's hoping for a correct answer. I'm just a researcher who wants to understand how you're making sense of these problems. Uh again, read the question aloud to me and describe aloud what your thinking as you're answering the question, and I may remind you to keep talking to keep the conversation moving.

Pre-questions



Hospital. A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50% of all babies are boys. The exact percentage of baby boys, however, varies from day to day. Sometimes it may be higher than 50%, sometimes lower.

For a period of 1 year, each hospital recorded the days on which more than 60% of the babies born were boys. Which hospital do you think recorded more such days? Explain your reasoning,

A. The larger hospital

B. The smaller hospital

[✓ C.] About the same

- over a year ~~more~~ each ~~days~~ hospital is likely to have ~~two~~ the same ~~time~~ / b/c there is a lot of data, compared to only collecting it for a few days.

S:
[00:07:10] Okay. So it says, a certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50% of all babies are boys. The exact percentage of baby boys, however, varies from day to day. Sometimes it may be higher than 50%, sometimes lower. For a period of one year each hospital recorded the days on which [00:07:36] more than 60% of the babies born were boys. Which hospital do you think recorded more such days? Explain your reasoning.

S: Um, I would probably -- I'm kinda -- like, I think... if it's over a year, um, you know the -- the larger hospital in a shorter period of time, might experience more s-- more than 60%, um, babies born were, like, boys. But over a [OK] year, I feel like they would be probably around the same, just because [00:08:13] -- it's like, when you flip a coin, if you flip it ov-- only ten times, then it's gonna usually it probably will vary, but if you do it, like, a hundred times, or 300 times, it's probably gonna be more than, like, 50% each way. [OK] So that's probably why I would say the -- about the same.

I: Okay. Um, so could you -- I just want to ask a [mh] couple questions about that. So you said something about the -- at the larger [00:08:39] hospital [mh] -- um, if it was a shorter number of days, you'd expect there to be more days -- you said something about something [Yeah] happening at the larger hospital.

S: Mh, now that I think it -- about it again, I probably would switch it around, just because the smaller hospital, like, the more -- like one or two babies can make a much bigger difference in the percentage, [OK] so I would probably say that in the smaller hospital you might find [00:09:00] more times where it gets higher than six -- higher than 60%, just because there's fewer -- a fewer number. [OK] Um, so it's just like looking at like the percentages, um, it's a lot more likely to vary.

I: Okay. Um... and -- um, okay. And so why is it more likely -- why is the smaller hospital more likely to vary?

S: Um, so if you -- so if there's only 15 -- about 15 babies born per day, if you add [00:09:35] one, that like s-- that percentage, um, of, like, out of 15, [mh] is a -- like going to increase or decrease a lot more than -- so if you add 1 by 45, it's kind of like [Hm], well, if you have like a hundred people, and only 2 people don't show up, you got 80% of the class showing up, or something. [OK] And then if you have only, like, 10 people, and then, like, 2 -- or like, if you only have like 15 people and two people show up, or [00:09:59] something like that, like, it's gonna be a much bigger difference.

I: Okay. And so what -- um, what answer, uh, were you choosing?

S: Um, about the same.

I: Okay. And so -- so the kind of the second part of your explanation had to do with, um, over the course of a year, you would no longer expect to see that difference. [mh] Could you say a little bit more about that?

S: Yeah, over just an extended period, I mean that's 365 days [00:10:27], so you get a lot more -- even if you the variations, usually it'll be around -- like, in total it'll probably around 50 uh, percent, just like when you flip a coin. [OK] Um, it's the same kind of percentages with that. Like, if you only do it, you know, ten times it's gonna be a little bit more varied, like, you might get, like, 70% of heads instead of more like 50. But if you do it like 100 or 300 times then you're more [00:10:56] likely to get that -- around that 50 percent.

I: Okay. Okay, great.

And can you put a couple bullet points [mh] describing your reasoning. I'm just gonna make sure the camera is rolling.

/Written:
Over a year each hospital is likely to have the same, b/c there is a lot of data, compared to only collecting it for a few days
/

S: Of course.

I: Okay, great. And, um, how confident are you in your answer?

S: I'd say pretty confident.

I: =Pretty confident.=

S: =Or, at least= confident. /c/

I: Confident.

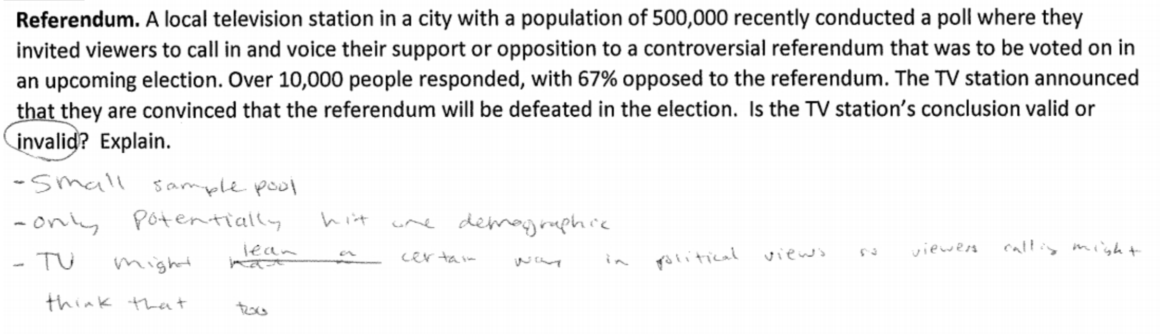
S: If not pretty confident. [/c/] In -- in between those.

I: Okay, sounds good. [00:12:20] And, um, can you repeat back to me the written problem?

S: So I was saying how there was a hos-- a bigger hospital and then the smaller hospital. Um, and about 50% of each, like, baby that's born is a boy. And then it asked, um, which hospital would be more likely to have, like, over 60%, over like a year I think it was. [mh] Um, and I said it was about even.

I: Okay, great.

** Referendum



Referendum. A local television station in a city with a population of 500,000 recently conducted a poll where they invited viewers to call in and voice their support or opposition to a controversial referendum that was to be voted on in an upcoming election. Over 10,000 people responded, with 67% opposed to the referendum. The TV station announced that they are convinced that the referendum will be defeated in the election. Is the TV station's conclusion valid or [✓ invalid]? Explain.

-Small sample pool
-only potentially hit one demographic
- TV might ~~have~~ lean a certain way in political views so viewers calling might think that too

S: Okay. A local television station in a [00:12:58] city with a population of 500,000 recently conducted a poll where there -- where they invited viewers to call in and voice their support or opposition to a controversial referendum that was what -- that was to be voted on in an ac-- in an upcoming election. Over 10,000 people responded with 67 opposed to the referendum. The TV station announced that they are convinced that the referendum [00:13:22] will be defeated in the election. Is the TV station's conclusion valid or invalid? Explain.

S: Um, I would say that even though on-- so only ten -- ten thousand responded out of, say, like, fif-- five hundred thousand of the, um, population. Um, but also if it's a TV station, usually, there're gonna to be more politically incline-- like, their viewers are going to be usually one side of the political era -- area. Um, and especially with [00:13:57] people calling in, anything can really change their, um, opinions. And some people that don't call in, um, there could be a large population of them that actually, like, do -- like, are going to be voting, but they don't want to call in, or they're working [Hm] or something like that. Um, for example, like, you look at the last election, things were pointing towards Hillary and then a ton of people decided [00:14:18] to vote Trump. Like, and there really wasn't -- it was going back and forth, like, each news station was kind of going like up and down with it. So I think especially with elections, um, you can't really just focus on one area, that's only one sample that they took. Um, and it might be some people that are more, a-- like, okay with calling in, or okay with voicing their opinions, but there's [00:14:40] also another [OK] s-- another part where there's people that aren't going to be more inclined to do that. So I would probably say that, um, the TV station's conclusion is probably, um, I w-- I guess, invalid. I -- or at least it's not entirely accurate, and it's not looking at the whole entire picture.

I: Okay. Um, what could you -- is there something you could do to change or improve the study [00:15:11] if you were at this TV station, or -- [Yeah] a better -- a better way that you could conduct the study?

S: Yeah, I mean I would probably look at the demographics, first. Especially the people that were, um, calling in. See what -- what they -- what their demographics were. Because if they're from on s-- like, spot in the city [Hm], the other side the city might have a completely different opinion on the [00:15:34] matter. So I would probably try to, um, not only have people call in, but also have some of the people at the TV station calling on and asking for their opinion if they wanted to give it. Um... [OK] and also, just -- I mean you can't hit ev-- like five hundred thousand, but probably maybe trying to get the sample size a little bit bigger.

I: Okay. And how big would you be looking at, maybe?

S: Mm, I would probably [00:16:04] say at least, like -- I'd say trying to get like more like at least fifty thousand? I mean, it's not a ton compared to the population, but it's still much better than only ten thousand. [OK] Um, ideally it would be much higher than that, uh, but like, twenty five percent, but in reality, most likely, you aren't going to be able to get, um, a TV station being able to go into-- through, like, every part of the city and [00:16:33] finding [Hm] that many people. So I'd say probably just shooting to get at least -- um, probably fifty thousand, maybe only twenty thousand, but...

I: Okay. Great. Um, so maybe just circle sort of valid or invalid. [OK] And then just, again, a couple bullet points. Okay. And how confident are you in your response?

/Written:
small sample pool
only potentially hit one demographic
TV might lean a certain way in political views so viewers calling might think that way too
/

S: Confident.

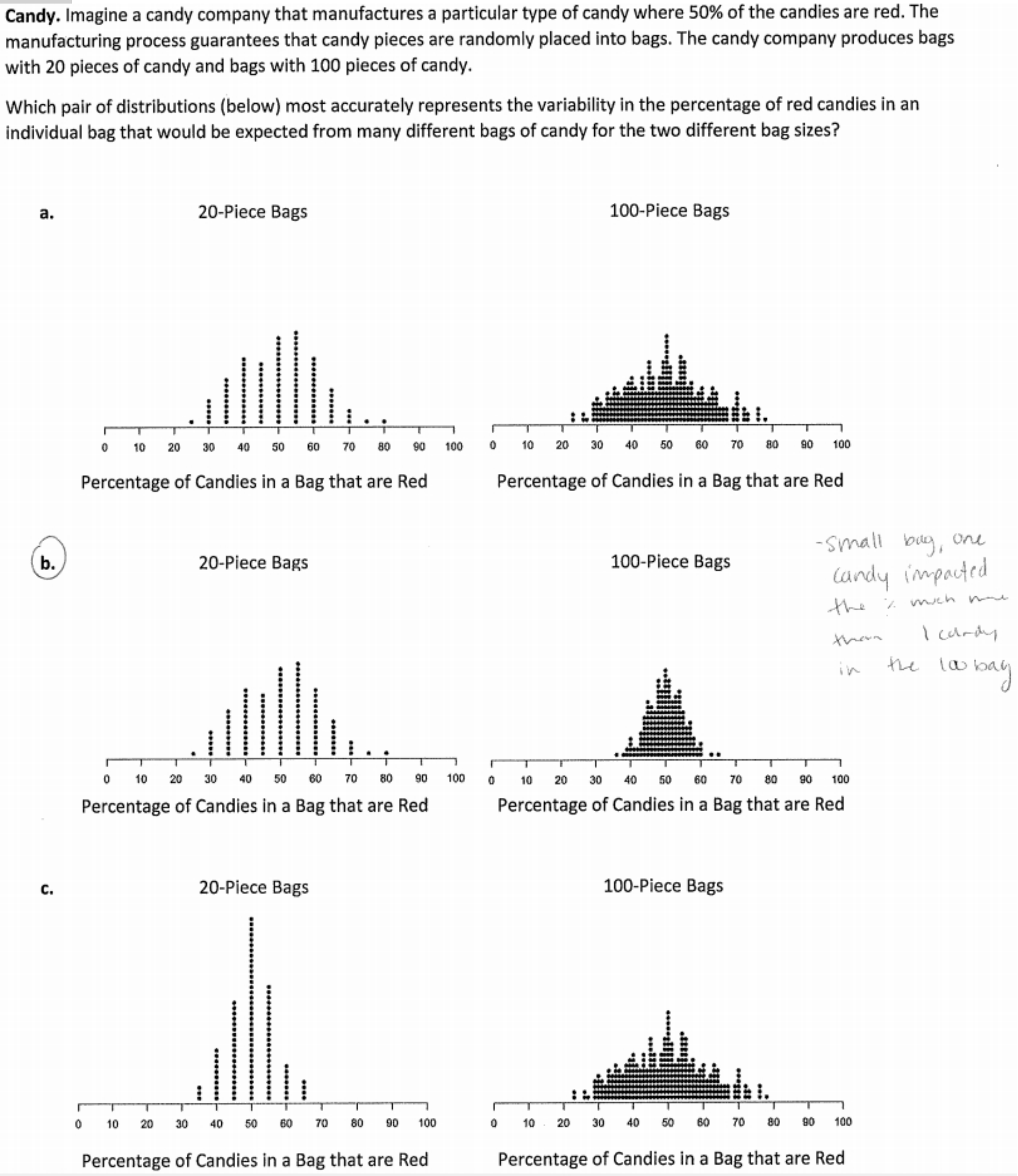
I: Okay. And can you repeat back to me the written problem.

S: [00:18:04] Yeah, so it was talking about how it was -- like a city was going through -- had a referending -- referendum that came up. And the te-- one of the TV stations, um, had, like, let people call in, and voice their opinions about it. Um, and they got about ten thousand people responding with 67% of um, them saying that they don't want the referendum to pass. And the question was asking, um, whether or not, like, you thought [00:18:31] the, um, the results I had were valid or invalid.

I: OK.

** Candy

1.460:



Candy. Imagine a candy company that manufactures a particular type of candy where 50% of the candies are red. The manufacturing process guarantees that candy pieces are randomly placed into bags. The candy company produces bags with 20 pieces of candy and bags with 100 pieces of candy.

Which pair of distributions (below) most accurately represents the variability in the percentage of red candies in an individual bag that would be expected from many different bags of candy for the two different bag sizes?

a. 20-Piece Bags | 100-Piece Bags

Percentage of Candies in a Bag that are Red | Percentage of Candies in a Bag that are Red

[✓ **b.**] 20-Piece Bags | 100-Piece Bags

Percentage of Candies in a Bag that are Red | Percentage of Candies in a Bag that are Red

-small bag, one
candy impacted
the % much more
than 1 candy
in the 100 bag

c. 20-Piece Bags | 100-Piece Bags

Percentage of Candies in a Bag that are Red | Percentage of Candies in a Bag that are Red

1.461:
1.462: S: OK. Imagine a candy company that
1.463: manufactures a particular type of candy
1.464: where 50% of the candies are red. The

manufacturing process guarantees is that candy pieces are randomly placed into bags. The candy company produces bags with 20 pieces of candy and bags with t-- 100 pieces of candy. Which pair of distributions, below, most accurately [00:19:03] represents the variability in the percentage a-- the percentage of red candies in an individual bag that would be expected for many different bags of candy for the two different bag sizes.

S: OK. S--

I: So what are you noticing?

S: So with the graphs, I can definitely see that certain -- depending on which one it is, um, they either have a much bigger range or much smaller range [mh]. Um, [00:19:40] and fifty percent of the candies are red, 20 pieces of candy in bags. 100 pieces of candy. Um, I would... probably... say... let's see. I would probably say that B would be most like what I would think. Um, [OK] just kind of going on one of my previous, like, responses. If you only have 20 pieces, it's a lot -- you're gonna have a much bigger range, because if you take one of [00:20:25] those away, it's going to change the percentage, um, a lot more /inaudible/ and than, like, if you like have a much bigger, um, bag of candies. So probably the 20 pieces of candy -- or the 20 piece bags for the percentage of candies that are red, like, it's gonna be a much bigger range, um, and then the 100 piece bags, um, having a much smaller range, just because if you take a couple out, um, it's not going to be -- it's not [00:20:52] gonna have as much of an impact on that, um, ratio. [mh] So.

I: Okay. Okay.

S: And then, do you want me to write?

I: Yeah, yeah.

/Written:
small bag, one candy impacted the % much more than 1 candy in the 100 bag
/

S: Okay.

I: Okay, and how confident are you in your response?

S: I'd say... confi...dent. But ehhh.

I: Okay. /cs/

I: Can you tell me a little bit [Just--] more about that?

S: Just, um, it kind of depends, also. Like, I -- I don't know. Just, like, how accurate, or if like there's [00:21:45] different weights for them, like, within, like, the manufacturing company. Like, um, sometimes, like, you know they might have -- have like a bigger wave coming in. Like on -- if they had like a track or something. [Hm] Like, depending on, like, when they come down and whatnot. [Hm] So, I think it's kind of depends on like just like the timing of it all. Like, if it's very accurate, and they try to really get, like, that 50%, like, that's [00:22:08] one [Hm] of their main goals [Hm] rather than, um, if they just want to get that 20 pieces in there, and they don't really care. Because it's kind of like when you open, like, a bag of Skittles, like sometimes there's no purple ones, or there's no red ones and like you're -- you're kind of like, "What?" Or, like, you open, like, a fruit, like, a snack and I'm thinking, like, the Scooby-Doo ones, like, I always want to get the blue ones [Hm] but [00:22:27] sometimes there's not even any in there, and sometimes there's five. [mh] So it kind of -- it kind of just depends, I guess.

I: Mh. Um -- so if that's saying that they're placing them randomly, [mh] um, and 50% of them overall -- can you say a little bit more about what you were saying about the timing? The timing of when they come down... yeah.

S: Yeah, kind of like, if say if they're in, like, two different buckets, like, if they're [00:22:58] being poured at the same time [Hm], like, onto the conveyor track, or something like that [Hm], um, then it'd be a lot more likely, like, they're gonna be, like, it's gonna be more like an even distribution like a ran-- like, it's gonna be like more, like, probably fifty-fifty or something, like, [mh] like, random. But if they're going at different times, you know, they're gonna have less time to, like, intermingle with [00:23:17] each other and kind of [Hm] mix up. Um, and it's kind of gonna be -- it might be, like, just cutting it off and then all of them might be, like, one color, or there might be only one or two [Hm] of the other color, so.

I: And how -- does that affect your reasoning about what these plots look like at all?

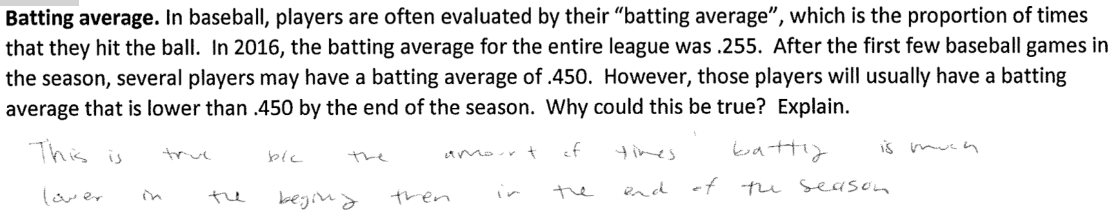
S: Um, I think I tried to simply it -- it's -- make it simpler just for this. [Hm] Just thinking about, like, most likely they're probably [00:23:42] going in at the same time or they're mixing them up and then [OK], like, putting them in. So I tried to, like, only really think about, like, making it -- trying to make it as simple as I could [mh], and then -- but there's definitely, like, other things that come into play, but. [mh]

I: OK, great. So, um, can you repeat back to me the problem?

S: Yeah, so it was -- there was, like, two types of -- or two bags of candies. Like one smaller one, and [00:24:08] one bigger one. And 50% of the candies were red. Um, and then it asked which graph do you think most represented the percentage of, um, the red candies in each bag.

I: Okay, great.

** Batting averages



Batting average. In baseball, players are often evaluated by their "batting average", which is the proportion of times that they hit the ball. In 2016, the batting average for the entire league was .255. After the first few baseball games in the season, several players may have a batting average of .450. However, those players will usually have a batting average that is lower than .450 by the end of the season. Why could this be true? Explain.

This is true b/c the amount of times battig is much lower in the begining then in the end of the season

S: Okay. In baseball players are often evaluated by their batting average, which is the proportion of times that they hit the -- [00:24:35] which is the proportion of times that they hit the ball. In two hun-- in 2016 the batting average for the entire league was 0.255. After the first few baseball games in the season, several players may have been -- may have -- may have a batting average of 0.45. However those players will usually have a batting average that is lower than 0.5 by the end of the season. Why could this [00:25:00] be true? Explain.

S: So just looking at how many times they probably hit the ball, um, it's going to affect it. Like, if they get -- the first two times they get-- like, in the -- their first, like, couple of games, if they keep hitting, like, home runs, or they keep hitting [Hm] it, and they get quite a few, like, ones out of it, um, it's gonna be a much higher batting average [mh], than if, say, out of like a hundred games, like, you know [00:25:24] they might have a few games where they just have off games and they just don't, like, really can't hit anything. [mh] Um, so it's more of -- and it's -- and it's the average, so, if you only have, like, say, like, four -- like, four or five, like, times that you hit -- try to hit the ball, it's gonna be more likely, like, it's, like, oh yeah I hit it or I didn't hit it. Rather than if I -- you try to do it 100 times, it's gonna be a much [00:25:50] smaller area, where one thing is, like, it's gonna be probably much lower, because the baseballs are very small so it's very difficult to [Hm] hit all of them, and hit them every single time. [mh] So.

I: And so, um -- and so what's the difference between, say near the beginning of the season [mh] and at the end of the season, or what's -- what's the...

S: Yeah. It's -- basically, like, just how many you -- times, like, you've been able [00:26:18] to go up to bat, um, [OK] it's gonna be much lower in the beginning of the season, because you probably have only played a few games, rather than at the end of the season, um, you're gonna be -- you probably have batted many more times and, um, probably missed quite a few more times than, say, [OK] in the beginning of the season. And -- so that's why it's going to be lower is that just on average, it's going to keep [00:26:40] on most likely /c/ getting lower instead of higher. [Hm] It's kind of like if you score, like, in a different sport, like, if you score once in one game and then you don't score the next game, you're gonna have a 50% scoring, [mh] like per -- per game. But in reality, like, it might be a very rare occasion where you score. [Hm] And then if you take, like, out of a hundred games, maybe you only score, like, three times in 100 games, so. [mh]

/Written: This is true b/c the amount of times batting is much lower in the beging then in the end of the season/

I: Okay, great. [00:27:36] And how confident are you in your response?

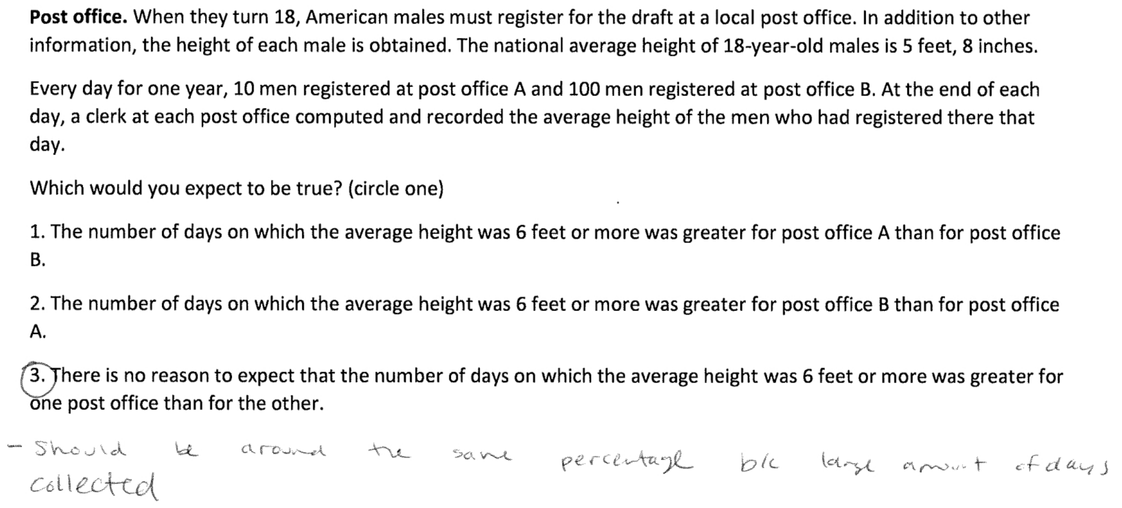
S: I'd say confident.

I: Okay. And could you repeat back to me the written problem?

S: So I was talking about the batting averages and, um, baseballs, so usually, um -- well over the whole entire league, it's point two five, I think it said, or two five five. Um, and then -- but in the beginning, a lot of times it's -- I think it said point four five. But as the season goes on, it's usually much low-- it usually [00:28:06] decreases, um, and it asked me, that -- is this probably true, or, if -- and then explain your answer.

I: Okay.

** Post Office



Post office. When they turn 18, American males must register for the draft at a local post office. In addition to other information, the height of each male is obtained. The national average height of 18-year-old males is 5 feet, 8 inches.

Every day for one year, 10 men registered at post office A and 100 men registered at post office B. At the end of each day, a clerk at each post office computed and recorded the average height of the men who had registered there that day.

Which would you expect to be true? (circle one)

1. The number of days on which the average height was 6 feet or more was greater for post office A than for post office B.

2. The number of days on which the average height was 6 feet or more was greater for post office B than for post office A.

[✓ 3.] There is no reason to expect that the number of days on which the average height was 6 feet or more was greater for one post office than for the other.

- Should be around the same percentage b/c large amount of days collected

S: Okay. So, um, the post office -- when they turn 18, American males must register for the draft at a local post office. In addition to other infor-- information, the height of each male is obtained. The national average of an eighteen year -- of eighteen year old males is five feet eight inches. Every day for [00:28:42] one year, ten men registered at post office A and a hundred men registered at post office B. At the end of each day, a clerk at each p-- post office computed and recorded the average height of the men who had registered there that day. Which would you expect to be true? The number of days which the average height was six feet or more was greater for post office A than post office B. The number of days [00:29:07] on which the average height for six -- six -- was six feet or more was greater for post office B than for post office A. Um, there was no reason to expect the d-- number of days on which the average height was six feet or more, um, was greater for one post office than for the other.

S: So it was every day for one year. So if it's -- if each office, like, depending on if it's, like, a n-- like, next -- cities next to it [Hm], like, um, [00:29:34] if they both have, like, the city average is about the same, I would probably say that, um, like, it was -- there's no reason to expect either one to be higher than the other, um, just because if they both have around the same average, then you can definitely compare them. I know, like, where I'm from [Hm], the -- everybody's really tall. [Hm] And so compared to, say, if I went, like, out east [00:30:01], like, I'd be more likely to be the same height there than I would be, like, from where I'm from. Like, I was kind of, like, the -- one of the shortest people there. And I'm [Hm] technically I think I'm average for height, n-- nationally. Um, so definitely like, hmm. Depending on what area you're from, so if they're -- say, like, they're both even, I would say there's no reason to expect, um, one to be higher than the other, but [00:30:24] even with, like, the bigger or smaller sample sizes, um, I think it really depends on the city's average, and, um, going from there. [OK] So.

I: Um, so -- so if we say -- if we say that they're, say, from the same city [mh], or from comparable cities. What would you -- where would you go from there?

S: I would probably say that there -- um, there's no reason to expect that the number of days on which the average height was six feet [00:30:52] or more was greater at one post office than another, because over a hundred -- or over a year, it should be -- it's kind, like, when you flip a coin. Like, the more you do it, the more likely [mh] it's going to be, probably, um, 50/50, so.

I: Okay, great. And a couple bullet points again. And how confident are you in your response?

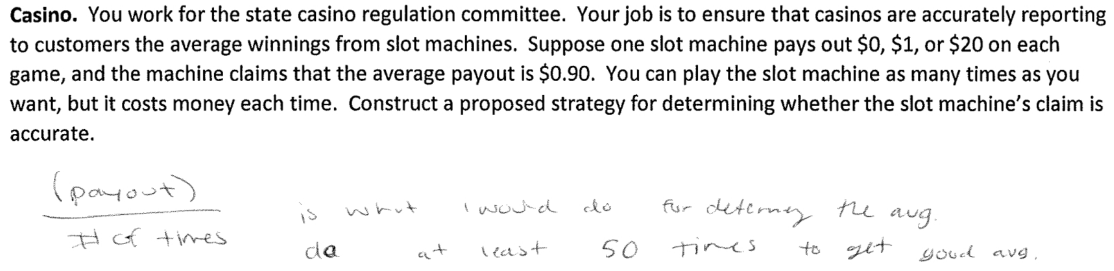
/Written:
Should be around the same percentage b/c large amount of days collected/

S: I'd say confident.

I: OK. Can you repeat back to me the written problem?

S: Yeah, so I was talking about, um, how [00:31:45] when men turn 18, they have to go to the post office and dr-- like, go, um, sign up for the draft. And they take the height, um, and then the national average is 5 foot 8 inches. And it was asking if a post office that got about 10 people each day, and then a post office that got out at 100 each day, um, for an entire year, they recorded all of the heights, and which post office would be more likely [00:32:09] to, um, have more of -- more peop-- more people like that are six foot or over, um, come in, like, a single day. /inaudible/ [OK]

** Casino



Casino. You work for the state casino regulation committee. Your job is to ensure that casinos are accurately reporting to customers the average winnings from slot machines. Suppose one slot machine pays out $0, $1, or $20 on each game, and the machine claims that the average payout is $0.90. You can play the slot machine as many times as you want, but it costs money each time. Construct a proposed strategy for determining whether the slot machine's claim is accurate.

$$\frac{(\text{payout})}{\#\text{ of times}}$$ is what I would do for determing the avg.
do at least 50 times to get good avg.

S: Casino. You work for the state casino regulation committee. Your job is to ensure that cons-- casinos are accurately reporting to customers the average winnings from slot machines. Support -- suppose one slot machine pays out zero [00:32:39] dollars, $1 or $20 on each game, and the machine claims that the average payout is ninety cents. You can pay the slot machine -- you can play the slot machine as many times as you want, but it costs money each time. Construct a proposed strategy for determining -- determining whether the slot machine's claim is accurate.

S: What I would probably do first is see how many people [00:33:03] play it -- pay -- play it each day, so say, like, you're gonna take data for, like, one certain day, um, I would probably see how many people are playing it. Also, how much it costs each time that you put money in, um, and what, like, the percentages are for, like, zero dollars, one dollar, or twenty dollars, um... [OK] Because if the twenty dol-- the twenty dollars is prob-- probably, like, very [00:33:28] rare, probably a low percent, um, but it is also, like, may happen quite a bit, because if it's one dollar, so they might be able to play it a couple times, or if there's any time where it's like, oh you get a free round, or something [Hm]. Um, just taking all the different percentages of, like, zero dollars, one dollar, twenty dollars, um, and then taking the price of each game and then how many people play [00:33:53] each day. [OK] So that's probably what I would do.

I: And what would you -- and feel free to [mh] scribble anything down if you want. Um, what would you do with all that information?

S: Yeah, so first I would probably add up all the payouts. So, like, the $0 $1 or $20. So I'd take all those numbers, put it on, like, say, like, a table, or put it into my [00:34:19] calculator, get the mean, or, like, the average of that number [OK]. Um -- or wait, let me think about this. Oh, wait, no. I wouldn't do that. I would just add up all the numbers /claps/ and then, um, I would probably subtract the average payout. Um -- or, no I wouldn't subtract the average payout. I would subtract the, um, money it costs for each time. So add up all of the payouts, so the 0 1 or 20, and then subtract how many, like, times people put the money

[00:34:52] in to actually play the game. [mh] And then divide that number by however many times people played it. [OK] So. And then that should give me the average payout. Um, so.

I: Okay. So, um, the average payout doesn't -- um,

S: -- =doesn't= --

I: =-- doesn't= take into account however much it costs.

S: Oh, OK.

I: So it's just whatever comes out of the machine.

S: Okay, so then I would just do the 0 1 or 20, and then divide by the number of [OK] people that play it.

I: OK. Um, so, um, I just want [00:35:28] to add one thing in here. So, um, the intention here is that, um, you don't necessarily have access to an existing -- data from existing customers. [mh] But you do have access to the machine itself. [OK] So you can pay -- you can play -- you can play the machine yourself. So it wouldn't be just gathering data on other people using it. Um, but of course it costs some money each time. [mh] Um, so based on that, if you were [00:35:59] just playing the machine, um, how would that -- as opposed to gathering data from other people -- does that affect what -- what you would do at all, or can you talk through that for me?

S: I don't think

it really affect, like, me playing it, or anybody. Because usually if you're at the casino you'd expect to pay money. Um, so if people want to keep on playing it, or stop playing it, they're gonna [00:36:24] probably pay whatever they need to, um, because they know, like, they don't really have another option, like /c/, they either play it or don't play it, so. [mh]

I: So, um, okay, so talk -- so can you just talk me through the strategy again, based on you playing it. Um.

S: Okay. Um, yeah, so for, like, the strategy for determining whether the slot machine's claim is accurate, um, like so I would take out however much money [00:36:53] that was made, so, um, like, the -- if I, like, one, say, -- $1 $2, or -- $1 or, like, zero dollars, or 20 dollars, or whatever the amount was that I got, um, and then dividing each -- that number by, um, whatever, like, times that I played it by. [OK] So.

I: And so you get to choose how many times you play it. So how many times would you want to play the game?

S: Hm. I probably -- depe-- it'd probably depend on how I -- while I did it the fir-- did the [00:37:28] first time. If I lost, like -- if I didn't make any -- if I lost money on the first time, I'd probably play it one more time, and then probably call it quits, like, at most, like, I'd play it four times, but otherwise I'd probably just do it like two times. So.

I: Okay. So you do -- you do -- it's not your own personal money. Um, so I think the idea is that the state -- the committee will pay for you to -- to -- to play it. Um, does that affect [00:37:59] your -- does that affect your =answer?=

S: =Might as well= don't stop, then, if it's not my money. [OK] Um, but I would probably at lea-- um, do it at least, like, say 15 times, or s-- [OK] like a -- a decent amount of times, like, where -- or maybe, like, even 50 times, um, just so that I could get, like, a good average. Cuz sometimes, you know, you might get tw--- like $20 or something and a coup-- like, three times in a row, you might get, like, only zero. You might not get [00:38:27] any money. So definitely trying to, like, do at least fifty, um, or more, [OK] I would say. Depending on, say, if it's, like, busy, and they're like, oh yeah, we -- we kind of want you to, like, come back a [Hm] little bit later. But definitely doing it at least probably fifteen -- or fifTY -- times, um, to try to get a good average instead of just, like, a one-time average, so.

I: Okay. Okay, great. Um, so yeah. You can put a few bullet points for that. [00:39:27] I'm sorry to go back in time.

/Written:
(Payout)/# of times is what I would do for determining the avg. do at least 50 times to get good avg.
/

S: No, you're good.

I: I just realized that, um -- I just realized to ask you to repeat back the -- the post-office problem. Um --

S: Oh, yeah.

I: And I know it's switching a little bit, but --

S: No, you're good. Yeah, so with the post office, um, people that were 18 -- like, men that were 18 years old, they had to go to the post office, register for the draft, and they took the height, and so it was talking about how, uh, the average height was [00:39:52] -- that was req-- like, for nationally, that's five eight, and one post office had ten people -- um, or te-- about ten people that registered for the draft and then the other post office, um, had about a hundred people coming in. And then, um, it was a-- and then they recorded at the height each time. And then it was asking, um, whether or not, like, there'd be a big difference between each of the heights, um, that were [OK] [00:40:15] like, say, like, over six foot, so.

I: Okay. Um, and so back to this problem. [mh] Um, so you just wrote some quick notes about this, I just wanted to ask how confident you are in your answer to this one?

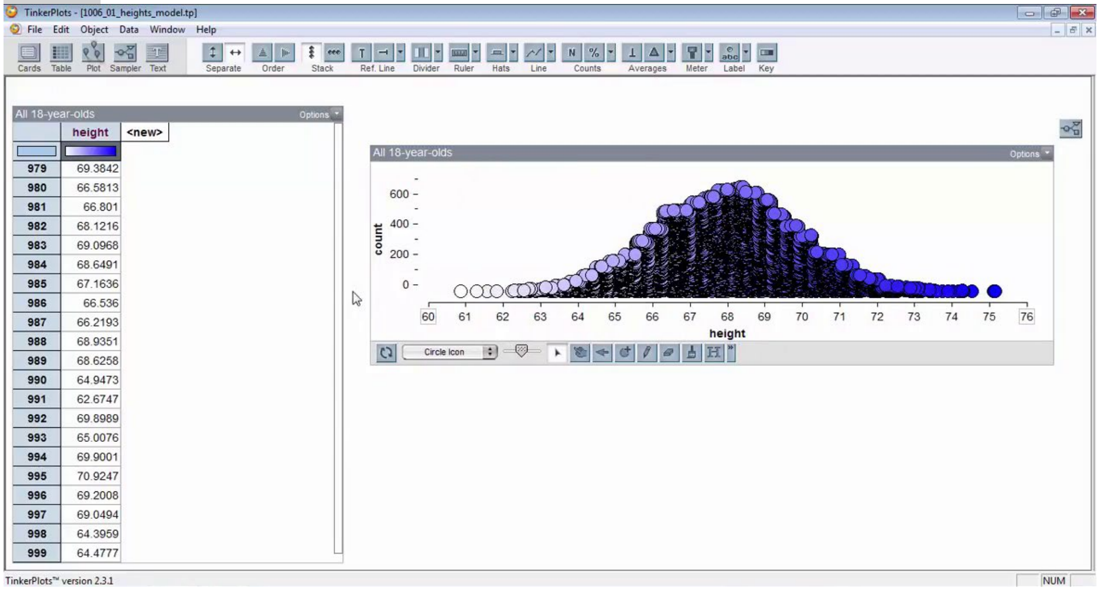
S: I'd say confident. Maybe. Yeah. [/c/] We'll say that. /c/

I: We'll say that. Okay. Could you tell me a little bit more about...

S: I can thi-- I guess it just kind of -- depending on, like, the machine, it [00:40:45] definitely, like, depends on what it feels like. I mean, it's random so it doesn't really -- it's gonna be -- it, like -- if the avera-- depending on, like, this average payout, [Hm] like, you -- you don't really know, like, what the, like -- whether they got that data, it's just basically, like, a number. They could have done it only, like, one time [Hm] and then said, oh this is the average pay-- mm -- like -- then like this is the average payout or [00:41:08] something. So, um -- but they could have done it, like, say, like, 10 times. And then ten -- or, like, nine times they got one dollar, one time they got zero dollars, and so it kind of put down the number, um. So you really don't know, like, how big of, like, set of data they were looking at when they took that average. [Hm] [mh] Um, so I guess that's kind of -- that. But.

Post Office Simulation

I: Okay. Great. So people sometimes have [00:41:38] trouble understanding problems of this form. So I'd like to provide some visuals to help you understand exactly what the question is asking. So, um, here's a data set of individual heights and weight -- of individual heights similar to those that we described in the problem. So you can imagine that these are the heights of eighteen-year-old men in the city that [00:42:07] has post office A and post office B. Um, so now I'd like you to plot these heights, um, to get a sense of how they're distributed. [OK] Great now can you tell me what the overall average height is.


TinkerPlots - [1006_01_heights_model.tp]
File Edit Object Data Window Help
Cards Table Plot Sampler Text
Separate Order Stack Ref. Line Divider Ruler Hats Line Counts Averages Meter Label Key
All 18-year-olds
Options
count
600
400
200
0
60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76
height
Circle Icon
TinkerPlots™ version 2.3.1
NUM

All 18-year-olds

	height	<new>
979	69.3842	
980	66.5813	
981	66.801	
982	68.1216	
983	69.0968	
984	68.6491	
985	67.1636	
986	66.536	
987	66.2193	
988	68.9351	
989	68.6258	
990	64.9473	
991	62.6747	
992	69.8989	
993	65.0076	
994	69.9001	
995	70.9247	
996	69.2008	
997	69.0494	
998	64.3959	
999	64.4777	

S: Um, so it was sixty seven point nine seven four two.

I: So the average here is the same as what's listed in the problem. Five feet eight inches is 68 inches. Um, what proportion of [00:42:51] the men in the town have a height greater than 68 inches, would you estimate.

S: Um, more than 5 8? Or --

I: More than -- yeah, 5 8. =68 inches.=

S: =So probably, like,= around, like, 50%, I guess. [OK] Wait... I'm trying to think, now. /c/ Let's see this is 100. Yeah, I'd say, like, more than would be like 50% [OK] about.

I: Um, and about how many are more than 72 inches?

S: Um, let's see. I guess that's, like -- it's a pretty low percentage, but do you want -- can I do the divider and everything =if I want? Or just, like, here --=

I: =Um, I just hear your eyeballing, first.=

S: [00:43:38] Um, let's see. There's not that many, so I would probably say, like, even though, like, on the screen it looks like 25, I'd probably say, more, like 15 or 10, just because there's a lot more people, especially like looking at all the numbers, like, there's probably like a thousand things, so I probably say, like, 10 or 15 percent over the 72.

I: Okay. And you said something about how there's a lot of -- maybe a lot [00:44:07] of people [Yeah]. Could you say more about that?

S: Yeah. Compared to like -- if there was like a hundred people, um, that were taken -- like, their data was taken, and then say, like, you could like almost individually count and also, like, one -- like I said before, like one person will make a much bigger difference or impact on the, like, [Hm] say the average or what percentage it is, um, compared to if it's a thousand people [00:44:30] it's kind of like if you lose a $1 bill and you only have ten dollars in your pocket, you're gonna b-- you're gonna -- it's you're gonna notice a lot easier than say if you have a thousand one dollar bills and you lose one [Hm], like it's gonna make -- be much less of a difference, like, if you look at, like, say the percentages in /inaudible/.

I: Okay. Great. Um, so you can actually look at the divider if you [00:44:50] want, just to see...

S: See, I haven't gone TinkerPlots for a little bit. Let's see if /c/ I can do it. Get close enough. Um, is it like percentage? Oh yeah. So that's at -- oh, I guess it's 2%. /c/ Much smaller. /cs/

I: Okay, so, um -- now imagine that we'll pick ten people randomly from the city. [mh] So you can just describe for me, how could you use TinkerPlots to randomly draw ten people from this group of people.

S: Yeah so, I would go to one of the samplers and then I would probably use, like, the -- um, probably, [00:45:30] like, the little ball one, where you have like multiple -- each ball [mh] like represents like a height. Um, so I would put all the heights into that, and then do, like, the -- um, it taking ten people each time, with -- and then I would do without replacement. So, like, each time it's not, like, that same height is not replaced.

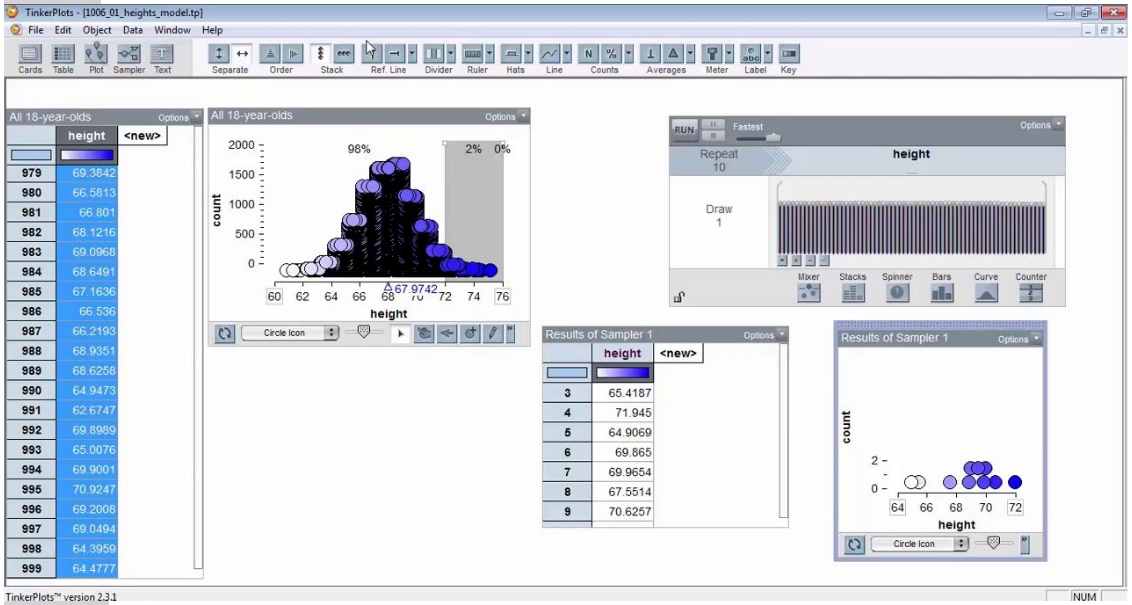
I: Okay. So, um, in fact, this file has all eighteen-year-old men in the city [00:45:57] in a TinkerPlots mixer, um, just like you -- just like you would do. [/c/] So each of these little balls is corresponding to one of those heights. And it's already set up without replacement, it's open on the top which indicates that. Um, so let's look at one sample of 10 people.

S: So should I change
the repeat to 10?

I: Sure.

S: Okay.
OK.

I:
[00:46:32] Um, can you plot that?
And, uh, what's the mean of that sample?



All 18-year-olds

	height	<new>
979	69.3842	
980	66.5813	
981	66.801	
982	68.1216	
983	69.0968	
984	68.6491	
985	67.1636	
986	66.536	
987	66.2193	
988	68.9351	
989	68.6258	
990	64.9473	
991	62.6747	
992	69.8989	
993	65.0076	
994	69.9001	
995	70.9247	
996	69.2008	
997	69.0494	
998	64.3959	
999	64.4777	

Results of Sampler 1

	height	<new>
3	65.4187	
4	71.945	
5	64.9069	
6	69.865	
7	69.9654	
8	67.5514	
9	70.6257	

S: Um, so it
was sixty eight point seven four.

I: Um, is the
mean gonna be the same every time we
draw a sample?

S: Probably not. It'll
pro-- it'll definitely vary depending
on, say, like, if you get -- s-- because it's
random, so you could get like two people
that are, like, only, like, sixty two inches, um,
and then a couple people like that or

[00:47:12] sixty eight which would definitely be -- make it quite a bit lower than what the actual average is of the entire sample. So it'll probably it'll vary a little bit, so.

I: Okay. Um, so the problem says that in post office A ten men registered every day for a year, and every day they noted the average height of the ten men. So can you set up a TinkerPlots model for taking the average of ten men and [00:47:35] recording that result?

S: So the ten people, and then...

I: So taking the average and then recording the result.

S: So taking... wait, can you repeat that?

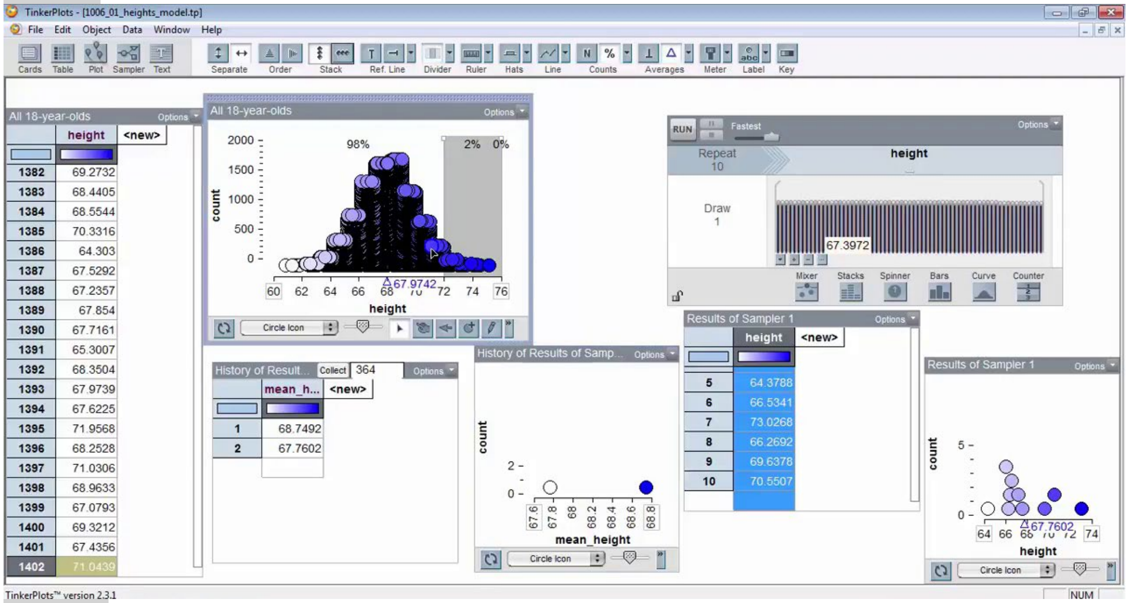
I: So, um, in the problem, um, they recorded how many -- they recorded, um, what their height was every day -- the average height every day for the year. So how would you set things up in TinkerPlots to record the heights every day for a year?

S: Um, I would [00:48:12] -- is it asking for, like, the mean each day.

I: Yeah.

S: Okay. So then I would do the record s-- or collect statistic [OK], and then I would do the three-- I would do 364. [OK]

I: Okay. So before we do that. Um, I'll just take one more sample. And, um, I'm gonna create a plot [Yeah] of this one. {Whoo. Thought that was gonna work. OK.} So, um, now we have three plots, here. Um, what [00:49:03] does the one dot in this plot represent. /population/


TinkerPlots - [1006_01_heights_model.tp]
File Edit Object Data Window Help
Cards Table Plot Sampler Text
Separate Order Stack Ref. Line Divider Ruler Hats Line Counts Averages Meter Label Key
All 18-year-olds
Options
count
2000
1500
1000
500
0
98%
2%
0%
60 62 64 66 68 70 72 74 76
67.9742
height
Circle Icon
RUN
Fastest
Repeat 10
Draw 1
67.3972
Mixer Stacks Spinner Bars Curve Counter
History of Results of Samp...
2
0
67.6 67.8 68 68.2 68.4 68.6 68.8
mean_height
Results of Sampler 1
5
0
64 66 68 70 72 74
67.7602
TinkerPlots™ version 2.3.1
NUM

All 18-year-olds

	height	<new>
1382	69.2732	
1383	68.4405	
1384	68.5544	
1385	70.3316	
1386	64.303	
1387	67.5292	
1388	67.2357	
1389	67.854	
1390	67.7161	
1391	65.3007	
1392	68.3504	
1393	67.9739	
1394	67.6225	
1395	71.9568	
1396	68.2528	
1397	71.0306	
1398	68.9633	
1399	67.0793	
1400	69.3212	
1401	67.4356	
1402	71.0439	

History of Result... Collect 364

	mean_h...	<new>
1	68.7492	
2	67.7602	

Results of Sampler 1

	height	<new>
5	64.3788	
6	66.5341	
7	73.0268	
8	66.2692	
9	69.6378	
10	70.5507	

S: So that represents, um, a height of an individual.

I: Okay. And what does one dot in this plot /sample/ represent?

S: So that's, um, in a group of 10 people. It's just the height of that one person from the sample of that [OK] -- the big one.

I: And what does one dot in this plot /ESD/ represent?

S: So that's the average of taking ten people and over the -- the entire [00:49:31] average. So it's having, like, say, like, the fit-- like, what is it -- however -- however many, like the tot-- all 100 or how -- all eighteen-year-olds, um, and then taking ten of those, and then the average of that.

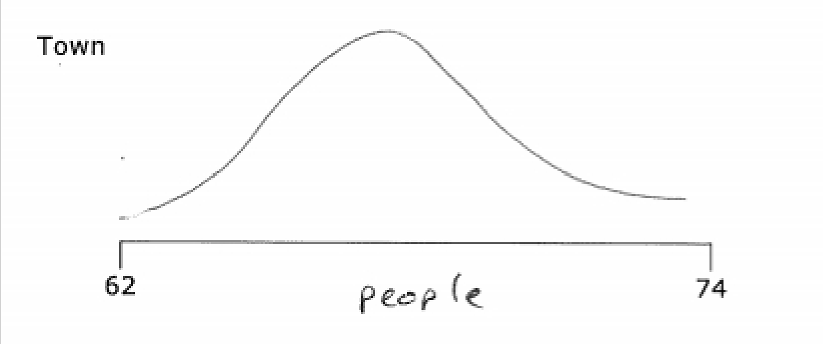
I: Okay. Um, so I've guided you through the TinkerPlots setup a bit here, but I'd like to return to my role as a researcher [mh], um, who's interested in how you're making sense of the situation. So imagine [00:49:56] that we repeat this process every day for a year, just like in the problem. What proportion of the sample means -- the dots in this graph of averages, would you expect to be over 72 inches?

S: The a-- the -- the average? =I'd probably have a--=

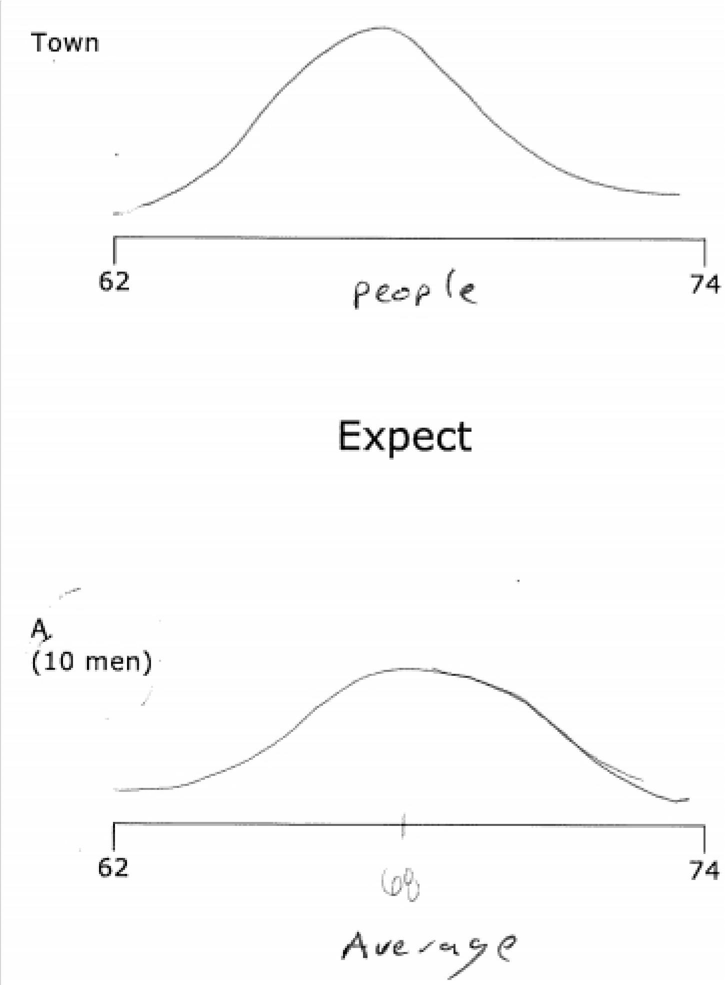
I: =So what proportion of= those -- these [Yeah] averages would be over 72 inches?

S: A very small proportion. [OK] Just cu-- I'd just say it's probably very unlikely, um, probably, like, maybe 1%. Because if you look at the [00:50:27] total height, um, over, like, however many people it is, it's only 2%. Um, and that's not that many, especially if you're taking a sample of ten. Like, most likely, you're not going to be getting, um, people that are over 72 inches. So I'd say it's a very low percentage. Like, maybe 1% of the peop-- like, a-- 1%, maybe, if the average would be 72, of, like, the averages.

1.1325: I: Okay. So -- um, so I'd like you to -- to
1.1326: [00:51:01] first sketch a graph of the town. And
1.1327: actually I'm gonna -- just so we're all on
1.1328: the same page, here I'm gonna change the
1.1329: axes here to go from 62 to 74. It's just
1.1330: gonna cut off a few values. Now,
1.1331: I'll do the same here.
1.1332: Um, so sketch the graph of the town, first. [OK]
1.1333:
1.1334:
1.1335: S: Can I just do, like, a curve?



1.1336:
1.1337: I: Yeah.
1.1338:   [OK] Just a curve. I won't ask you to draw 9,000 dots.
1.1339: And what do you expect to see for
1.1340: [00:51:49] post office A, when you have ten men -- the
1.1341: average of ten men. [Um...] So for this graph, here.
1.1342:
1.1343:
1.1344: S: Okay. I'd probably say, like -- so, like, this is 68, we'll say. Um, a little bit lower, but
1.1345: about there [OK] /c/.



I: Okay. And can you describe that graph for me?

S: It's just a, um, like a centered bell curve. Um, doesn't really sway one side or the -- to the other. [OK] Um, it's not skewed or anything, um, and just one -- one bump in the middle. [OK]

I: And how does it compare to the graph of the original town?

S: Um, it's similar, if not [00:52:31] about the same. I just had, like, a smaller amount of people.

I: Okay. Um, can you say that again. A smaller amount of people?

S: Like, it -- just, like, this is, like, um, the whole entire town, while this is only ten people.

I: Okay. OK. Um, and just to be clear, this is -- these are people -- and this is the average [mh].

I: Okay. Um, so let's see what actually happens. Um, so let's actually run this model. =So first--=

S: =Do want me to put --=

I: =Yeah, turn off, yeah, yeah.=

S: =-- make them all small.=

I: All that stuff. Yup.

S: We'll just make it all small so then [00:53:11] it doesn't -- it's the fastest...

I: You can leave that out.

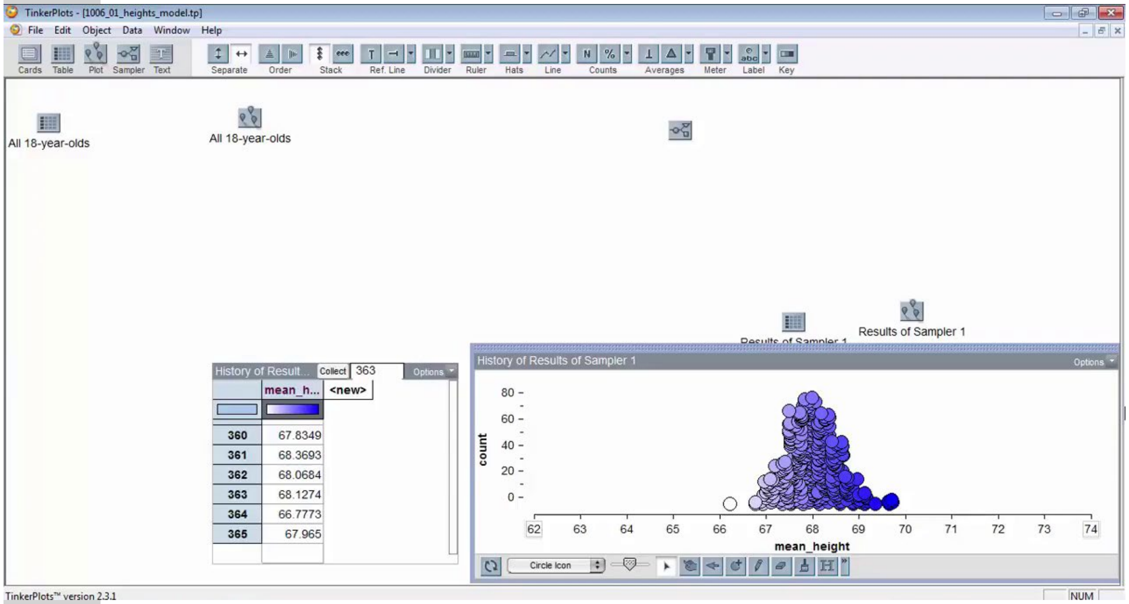
S: This one out? OK.

I: And then you want to turn off animation.

S: Uh... History Options. OK. Should I do this --

I: You can do it [Three.] three sixty three, yup.

1.1402:

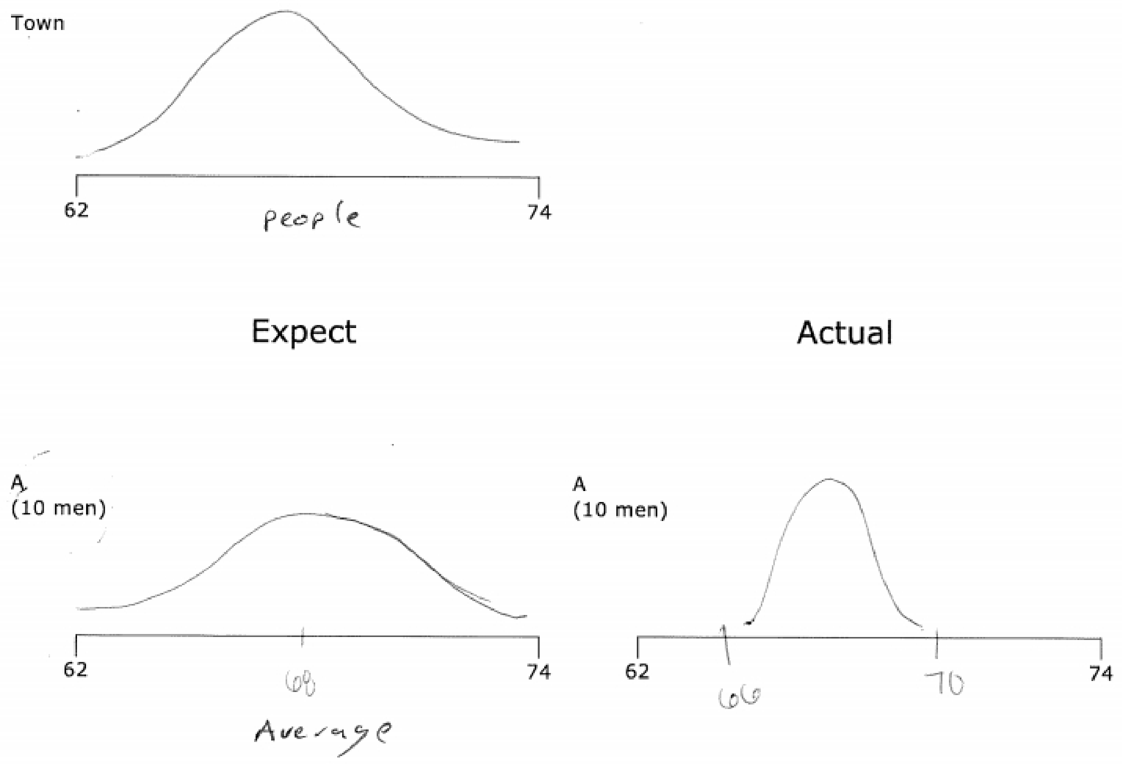

TinkerPlots - [1006_01_heights_model.tp]
File Edit Object Data Window Help
Cards Table Plot Sampler Text
Separate Order Stack Ref. Line Divider Ruler Hats Line Counts Averages Meter Label Key
All 18-year-olds
All 18-year-olds
Results of Sampler 1
Results of Sampler 1
History of Results of Sampler 1
Options
count
80
60
40
20
0
62 63 64 65 66 67 68 69 70 71 72 73 74
mean_height
Circle Icon
TinkerPlots™ version 2.3.1
NUM

History of Result... Collect 363 Options

	mean_h...	<new>
360	67.8349	
361	68.3693	
362	68.0684	
363	68.1274	
364	66.7773	
365	67.965	

1.1403:

1.1404:



1.1405:

S: OK. OK So, then, the actual, we'll put this as, like, 66. And then this is 70. So, basically, just goes like that, I would say. So it's a lot smaller of a range than what I had -- what I had said.

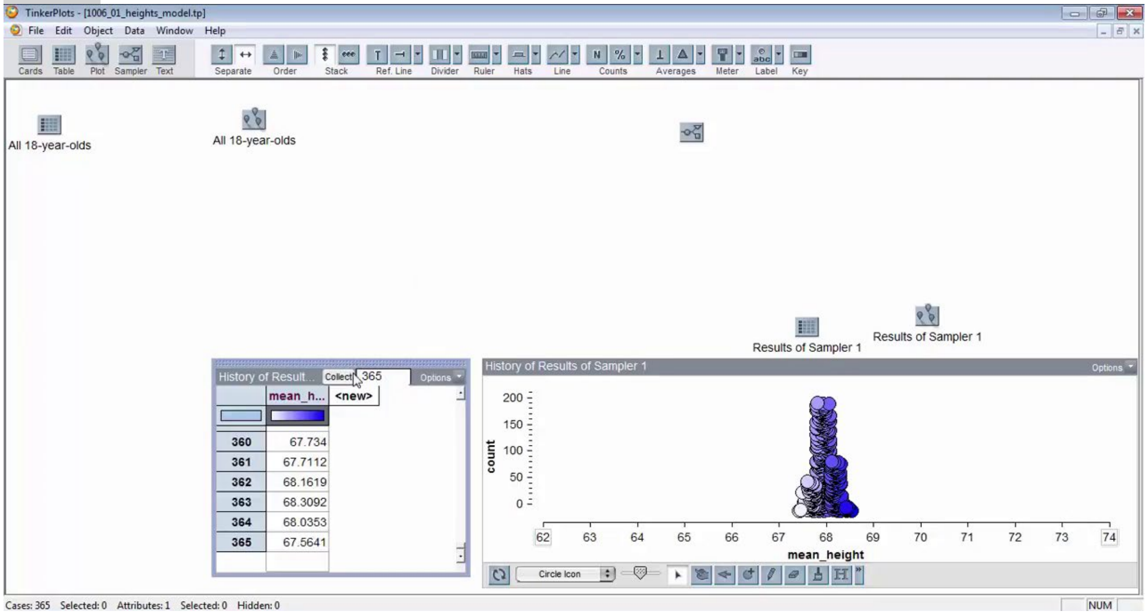
I: Okay. Is this surprising, or is [00:54:10] it what you expected?

S: No, it definitely makes more sense. Um, because most likely there's not gonna be a range of -- like an average of 74, if only like maybe one or two people got 74, there's no way I can really, like, be able to go all the way over there, um, if you take like ten people. Cuz, like, if you have, like, even just, like, one person at 74, you'll never be able to go to over to 74. [OK] So.

I: [00:54:33] OK. Um... Okay, so let's think about post office B, now, where a hundred people register each day. And so imagine that we generate, um, the same plot of averages, um, now for averages of a hundred, instead of ten. Um, what proportion of the sample means -- the dots in this graph of averages would you expect to be over 72 inches?

S: I'd say a very low percentage, again. Um, if that's [00:55:13] small -- or maybe -- maybe smaller? Um, than -- smaller than -- like, I'd say, probably around, maybe, one percent again, just, like, the ten men, um, may be slightly under one percent, just because you're taking much bigger, um,

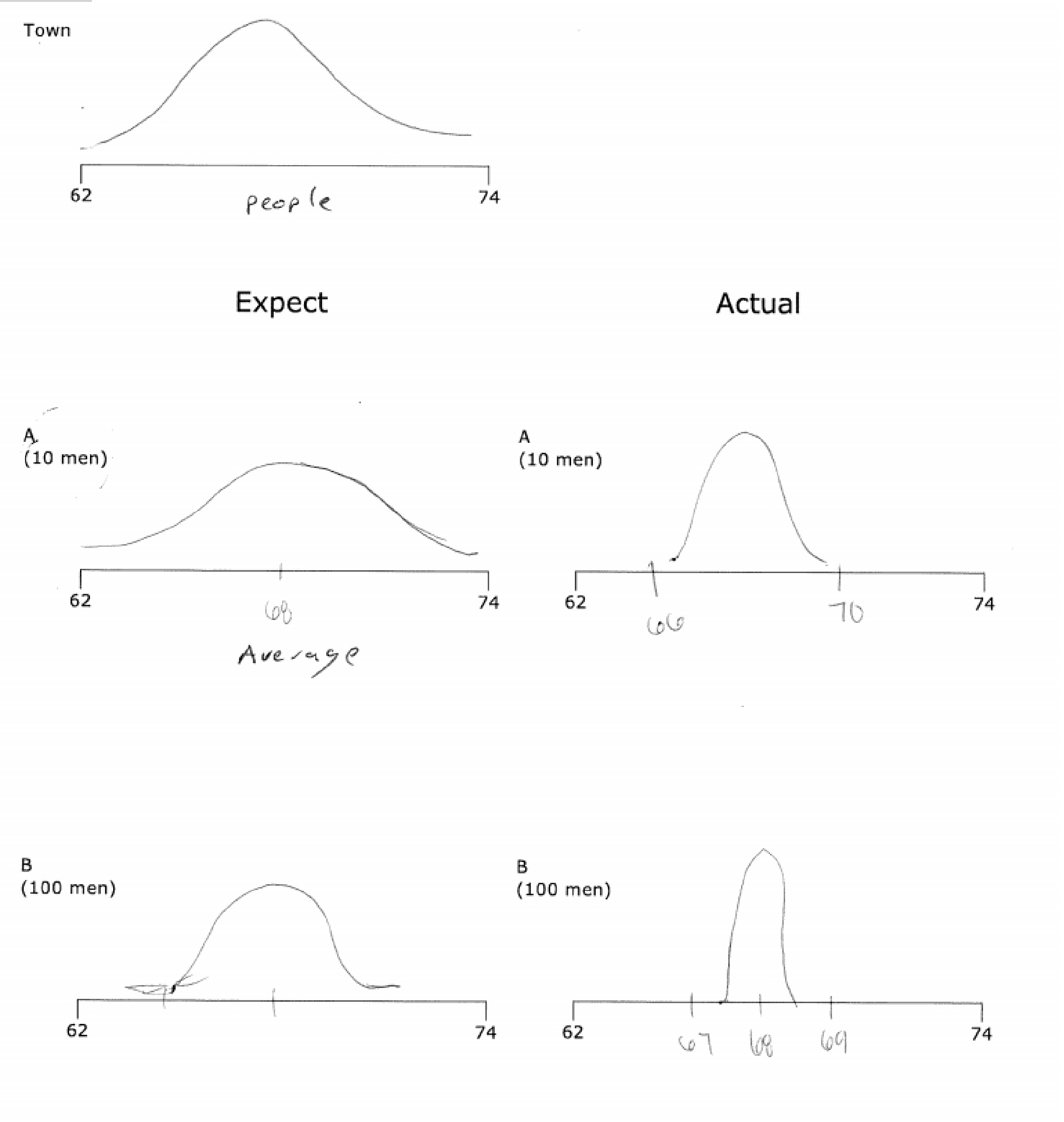
1.1447: amount of people, and most likely it's
1.1448: gonna be more towards that sixty-eight. [OK]
1.1449: But I would say, like, one percent or under.
1.1450:
1.1451:
1.1452: I: Okay, great. Um, so let's see what actually
1.1453: [00:55:42] happens. And I'll actually just quickly [Yeah] do this.
1.1454: [OK] Cuz we can -- I'm just gonna quickly
1.1455: change this... [/c/]
1.1456: once I convince it to open up from --
1.1457: just to 100. Shrink it right back down. Um, we
1.1458: can delete all history cases, just to
1.1459: clear that window. And then I'll just
1.1460: change this to 365 and we'll collect.
1.1461:
1.1462:


TinkerPlots - [1006_01_heights_model.tp]
File Edit Object Data Window Help
Cards Table Plot Sampler Text Separate Order Stack Ref. Line Divider Ruler Hats Line Counts Averages Meter Label Key
All 18-year-olds
All 18-year-olds
Results of Sampler 1
Results of Sampler 1
History of Result...
Collect
365
Options
History of Results of Sampler 1
Options
count
200
150
100
50
0
62
63
64
65
66
67
68
69
70
71
72
73
74
mean_height
Circle Icon
Cases: 365 Selected: 0 Attributes: 1 Selected: 0 Hidden: 0
NUM

	mean_h...	<new>
360	67.734	
361	67.7112	
362	68.1619	
363	68.3092	
364	68.0353	
365	67.5641	

S: So it has much smaller range.
1.1463:
1.1464: I: Okay.
1.1465:
1.1466: S: Okay.
1.1467: So I guess this one would be until 67 69, 68...
1.1468: [00:56:32] it's basically just /c/ and then if I would
1.1469: have done it before I would have
1.1470: probably said, like, what was it --

I: Oh, I forgot to ask you to sketch that. =I'm sorry.=



S:
=I probably would=
have just gone like that. [OK]
Not quite as much [OK], as, like, the one before,
but.

I: Okay.
Um, and so, what do you notice,
um..

S: The range shrunk. /c/ [OK] Just, um, especially
compared to that -- the ten men, um, before.
[00:57:00] It was about -- just under -- or just above sixty
six to about 70, and this time it was,
like, sixty seven and a half, about to 68 and
a half, so it was only about one, um, point
difference, for -- between, like, the mean heights
of a [OK] hundred people so.

I: And so, um, why -- why
are these two distributions different?

S: Um, they're different because you're
taking a much -- a smaller sample size. So one
person is gonna affect the whole entire
[00:57:33] sample a lot more, um, than, say, if you have a
hundred people, you know, instead of
having -- like, i-- it's just gonna affect it a
lot less. Um, so you might have, like, yes
you might have somebody that's 74 inches
tall, but if most of the people you take
are 68 [Hm], um, inches, then most likely, like,
your -- your mean is going to be 68,
especially over 365 days. [OK] So.

I: Okay. So that's actually, um, all we have for
[00:58:08] now.

S: /c/ Okay.

I: Um, any other questions or comments
before we close?

S: No.

Appendix D2: Growing Sample Proportions and Means

2.1: Growing Sample Proportions: Physical Simulation

2.2: I:
2.3: [00:00:02] So welcome back.
2.4:
2.5: S: Thank you.
2.6:
2.7: I: So today I
2.8: have a box with one, um, orange block, and one
2.9: blue block in the box. [OK] Um, we have a sheet here. And, um, we're
2.10: gonna shake the box, take out a block,
2.11: record it on this sheet, um, so O for orange,
2.12: and B for blue, and then I'll put the
2.13: block back into the box. Um, we're also gonna
2.14: mark how many blue we've seen so far and
2.15: the proportion blue that we seen so far. [OK] You can
2.16: calculate, um, using the calculator if you
2.17: [00:00:49] want to calculate anything. Um, and just as
2.18: an example, /inaudible/ our first one. So can't look at them when I pick. /cs/ This one was blue. [OK] So
2.19: we would B [B...] for blue.
2.20:
2.21: S: =And then, one, and then one, uh --=
2.22:
2.23: I: =And we've seen one
2.24: so far, and then it's hundred percent.=
2.25:
2.26: S: Should I
2.27: just do 100?
2.28:
2.29: I: Sure. Um, so before we continue,
2.30: um, what do you expect this table to look
2.31: like?
2.32:
2.33: S: Um, probably, like, just random. I
2.34: don't really know -- like, if it was -- it'll
2.35: be -- it -- it like might be close to 50, but also,
2.36: like, it could vary just because [OK], like,
2.37: [00:01:41] it's random, so.

I: Okay. So you'd expect the proportion blue at the end to be around 50 after 10 blocks?

S: If we did more trials than 100 -- or like than ten [Hm], but I'd say, like, maybe even, maybe more, maybe less. [OK] So, [OK] I can't really predict if it's gonna be random. But if it was a perfect scenario, then it would be like 50 per-- like, 50-50 [OK] but.

I: Gotcha. And what do you think will [00:02:11] happen sort of to the proportion blue like as it increases, like how is it going to change as it grows?

S: Um, I think it's gonna -- the proportion of blue compared to all the results will probably decrease, just because it's like, wow, it's one hundred perc-- like it's 1 out of 1 for, like, everything right now. But it's, like, you've only had one, so the more you put in, it's kind of like the -- the baseball example last time, how [00:02:35] like [OK] as the season goes on and they hit more, their, like, um, batting average usually decreases. So I would say that the proportion blue so far would decrease over the sample -- or, like, however many times you do it.

I: Okay. And would you expect any sort of patterns or anything in just the blocks, the B's and O's?

S: No. Probably not [OK]. /c/

I: Okay. All right. So let's see what
[00:03:00] happens. [OK] We'll just keep going here.
Um, blue.

S: I'll just put 100 at the very end here.

I: Sure. [OK]
Blue.
Blue. /cs/
Orange. Um, you can put -- the --

S: Percentages on?

I: Put the
percentages in.

S: OK. I'm gonna do this one here.
80 percent.

I: Orange. Orange.
[Oops.]
Orange. /cs/ Blue.

S: 5 out of 8.

I: Uh, 5 out of 9? =Oh, OK.= /c/

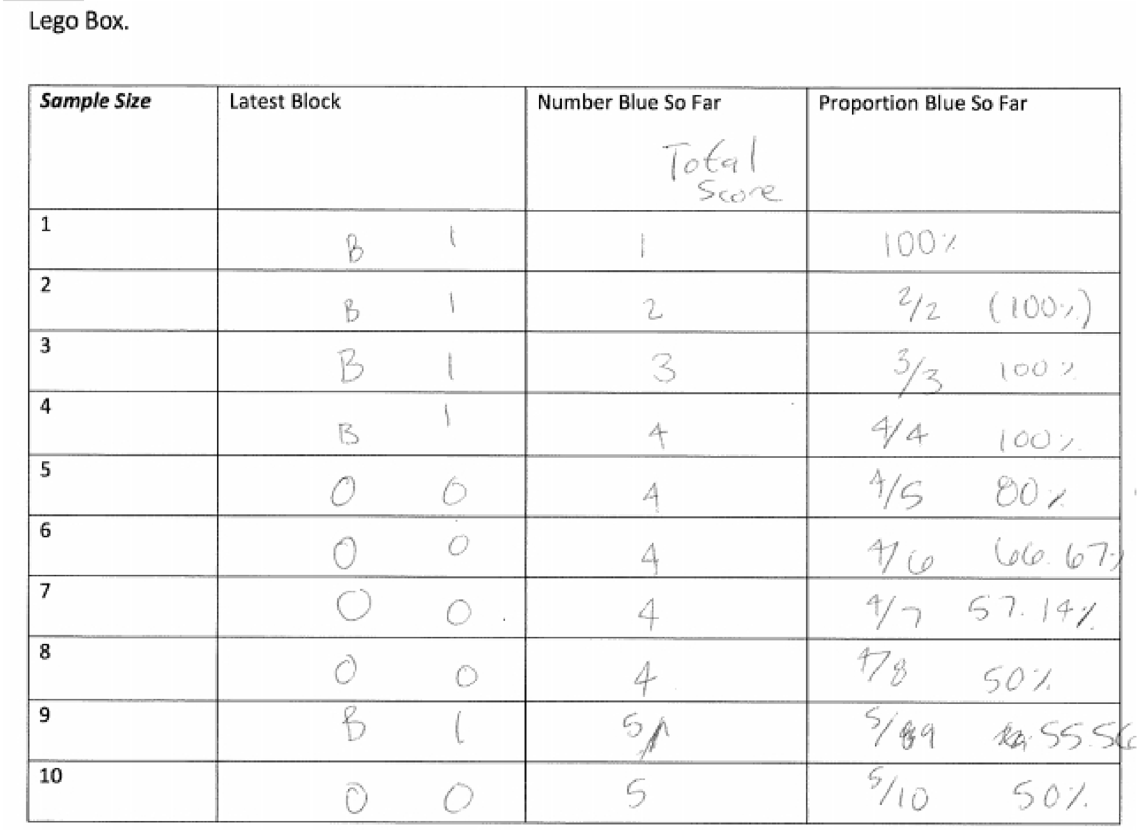
S: =Oh -- yeah. Yup. Nine.= I just can't do math.

I: So you can do that.

S: /inaudible/ All right.

I:
[00:05:12] Orange.

S: Five -- 5 out of 10. So 50%.



Lego Box.

Sample Size	Latest Block	Number Blue So Far Total Score	Proportion Blue So Far
1	B 1	1	100%
2	B 1	2	2/2 (100%)
3	B 1	3	3/3 100%
4	B 1	4	4/4 100%
5	O 0	4	4/5 80%
6	O 0	4	4/6 66.67%
7	O 0	4	4/7 57.14%
8	O 0	4	4/8 50%
9	B 1	5	5/~~8~~9 55.56
10	O 0	5	5/10 50%

I: So what do you notice about the proportion of blue blocks as we k-- keep on drawing more blocks?

S: It decreased. /c/ [OK] So. Yeah. Cuz it went from 100 -- t-- it went down about like 20 percent almost every time. [OK] Like, a-- I mean it -- as like we did more, it was less than 20 percent, [OK] but each time, like, there was, like, an orange block drawn, so.

I: Okay. And you said [00:05:49] something about it being less than -- something being less than 20 percent, or what were you =saying about twenty percent?=

S: =So, like, it got--= once we did like a couple, like, the first four were all blue blocks, and then when we got one orange, it changed it down from 100 to 80 [mh], and then from 80 to 66, so it's less than 20%, but -- and then it was like 66 to 57 [OK], so it just kind of kept going down, like, less, but [OK] it still [00:06:15] was a very drastic difference, so [OK] each time.

I: And so why do you think it kept going down less?

S: Because you have a bigger, like, sample size. So it's kind of, like, if you have like a wallet full of like $1000 -- like a thousand, like, one dollar bills and you take one out [mh], you're gonna probably forget about it [Hm] or, like, it really won't make a difference in the long run, while, like, if you only have $10 [00:06:40] and have like -- of $1 bills, and you have -- take one away, like, you're gonna have a lot -- like, you're not gonna be able to pay for a lot less. Or, like, a lot more and stuff like that. So. [OK] But, just in general, like, one out of like -- if you're like 999 dollars out of a thousand, like, the percent it's not really gonna change that much, [mh] just because of, like, how big of it -- how big it is [00:07:04], while, like, having 9 out of 10, that's, like, a much larger difference, so. [OK]

I: Um, what do you think would happen if we kept on drawing more blocks?

S: Um, I think it would probably stay around 50, just because -- um, you kind of have, like, a 50-50 chance. It'll probably go up and down, like, like, every time, like, we get like an orange or blue [mh], but if you did like, say, like 500 -- it 500 times, like it'd probably stay [00:07:35] around 50.

I: Um, and if we started over and did this again, what would you expect to be similar?

S: Um, probably that it would -- the proportion of the blue would either ch-- like, increase or decrease. Like, say, like, we had all oranges in the beginning [mh], like, it'd be 0%, and then we got blues, then, like, then it would go up. Or if, like -- if there's, like, a large portion, like, it'll go up or down. [mh] But it'll probably be just -- [00:08:09] at the very end it'll probably be maybe arou-- like, give or take, like, say, like, 10% for 50, just because of, like, how many, like, slots there are so.

I: Okay. And what would you expect to be different if we did it again?

S: Um, I prob-- maybe, like -- potentially, like, not as many, like, blue in the m-- like, how -- and how there's like two chunks, like, [Hm] because that was just, like, a random [00:08:37] occurrence. Um, like, maybe it'll happen again, maybe it won't, but... [OK]

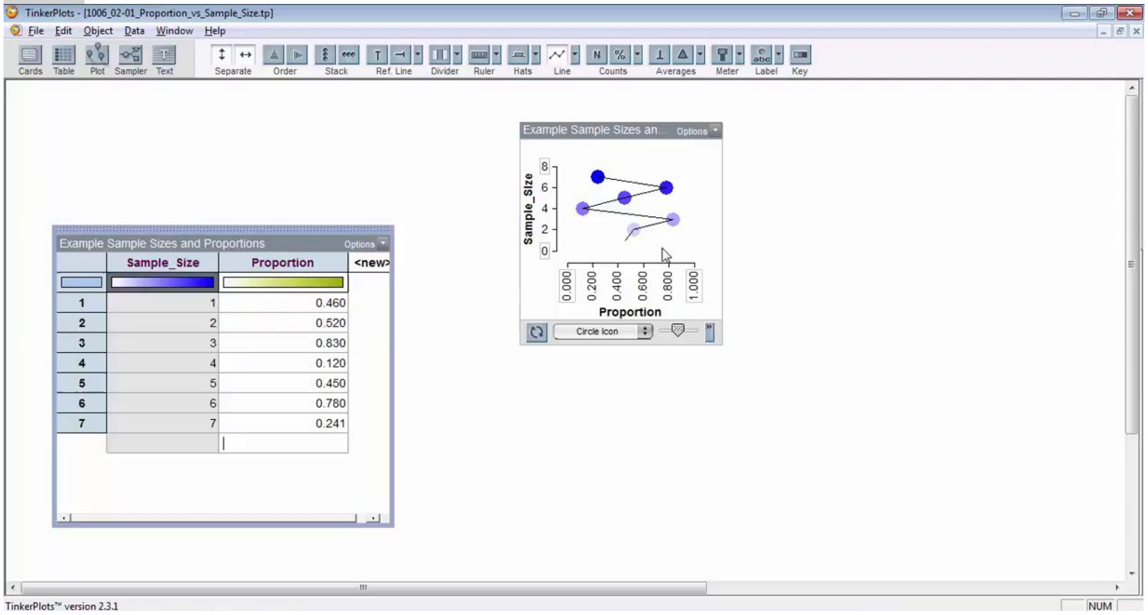
Growing Sample Proportions: TinkerPlots™

I: OK. Um, so, um, as we're kind of exploring how things grow, we're gonna be, um, using a new kind of plot which we can just call a sample size plot. Um, um, so this table shows a sample size and a proportion at that sample size, kind of similar to the table you created. Um, the -- these numbers were generated randomly just for this example, so they don't [OK] [00:09:12] represent anything in particular. Um, so let's create a plot of this data. Um, so I'd like you to plot the -- so drag down. And drag proportion to the x-axis. You can drag it out to separate the dots. And then drag sample size to the y-axis, and then drag it up to separate the dots. Um, and to kind of be able to see the trend as sample size increases, there's a couple just visual things we're going to [00:09:49] change in this plot. So first I'd like you to change the circle icon to borderless and then to add the line. Um, so the sample size plot shows, um, how the proportion changes as the sample size increases. Um, so, you know, for sample size two it was point five two and then it goes up to 0.83 and then it goes down to point 1 2. [mh] So, that's sort of what it's showing [mh] and the sample size is going this way. So the [OK] larger [00:10:23] sample size kind of goes up. So, um -- so I'm wondering what you think the sample size plot would look like for the proportion blue, which is what we were tracking here. I think we have it as a percentage but I have it as a proportion on this plot. So can you sketch what you think might happen, um, as you grow from 1 all the way to 50.

S: Yeah. Okay, so then -- so this would be, like, a hundred percent, like, [Yep] of, like, blue blocks?

I:
[00:11:00] Yep. Exactly.

2.258:
2.259:
2.260: S: Um, I'd probably say, like -- um...
2.261: I'm just, like, thinking about, like, if you
2.262: start with, like a -- like, a chunk, like, it's
2.263: gonna go either one way or the other.
2.264: Like, for starting wise. [OK] Um, so, like --
2.265: is it just, like, as you in-- I guess I could
2.266: do it this way. OK. I got it.
2.267: [OK] I think.
2.268:
2.269: I: Um, so, um, in this plot, so if I keep on
2.270: adding proportions on here. I'll just
2.271: [00:11:51] make up some numbers, 0.5...
2.272: I -- I'm just [mh] -- these are just random numbers
2.273: I'm just pulling out of m-- my head. Um, um, where
2.274: it's always going up. [Yeah] So there's a --
2.275: there's the proportion --
2.276:


TinkerPlots - [1006_02-01_Proportion_vs_Sample_Size.tp]
File Edit Object Data Window Help
Cards Table Plot Sampler Text Separate Order Stack Ref. Line Divider Ruler Hats Line Counts Averages Meter Label Key
Example Sample Sizes an... Options
Sample_Size
0
2
4
6
8
0.000
0.200
0.400
0.600
0.800
1.000
Proportion
Circle Icon
TinkerPlots™ version 2.3.1
NUM

Example Sample Sizes and Proportions

	Sample_Size	Proportion	<new>
1	1	0.460	
2	2	0.520	
3	3	0.830	
4	4	0.120	
5	5	0.450	
6	6	0.780	
7	7	0.241	

S: OK, yeah.
2.277:
2.278: I: -- at this -- at
2.279: 1, and then a 2, and then at 3. So you'll
2.280: need something kind of =going up= --
2.281:

S: =Like a zig-zaggy...=

I: -- from 1 or to 50, or whatever shape it is. [OK] Yep.

S: So, then, for, like, the actual I could just, like, draw, like, it --

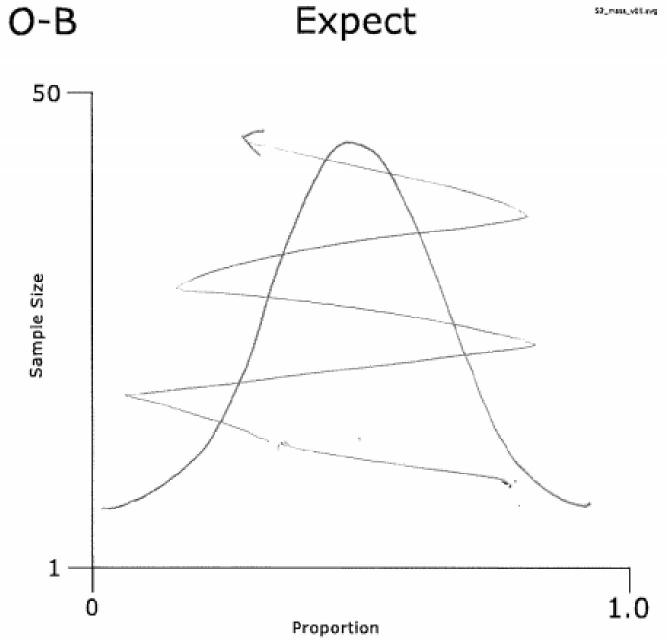
I: Um -- so can we -- yeah, can [00:12:23] you just draw -- you just draw over the =old one.=

S: =Over this= one?

I: Yeah, yeah, yeah, yeah.

S: Okay. So I guess, hmm, oh... /c/

I: Okay.



S: /inaudible/ there. [OK] /cs/

I: So it -- um, so can you describe what you just drew?

S: So, basically like it's alway-- the, um, sample size is always going up. [OK] Um, as you go on through, like, out the experiment. So you'll start, like, say, at, like 1 or whatever you start at. And then you'll just keep on increasing, so that's why it's always, like, going up, [00:12:59] and then, um, it, like, will go probably -- it'll kind of bounce between the two sides, like, like, 1 if it's like a blue or something, like, it could one side or the other, [OK] and it'll kind of go in between, like, um, probably, the, like, point -- for a proportion of 0.5, so it'll probably kind of stay around 0.5, maybe, and then just kind of go up.

I: Okay. OK. Cool. Um, so how can we, um, set up the block -- so I've [00:13:30] got a blank TinkerPlots file here. Um, how could we set up the block situation with the orange and the blue block in TinkerPlots?

S: So, like, having, like, a -- like the sampler, or... like making it -- like a sampler for this?

I: Sure. Yeah.

S: Okay. Well, you would -- you could do, like, probably -- I would probably just do, like, a spinner or something, where you have, like, a 50 for -- 50 chance of, like, having, like, a blue block [00:14:01] and an orange block, and then, um, probably -- then, like, once you plot it, like, collect statistics for how many blue blocks you get each time, um, and then you can kind of, just, I guess, like, increase -- or you can just, like, plot -- for, like, how like, the inj-- you can plot, like, how many, like, blue blocks were for each time and then kind of make the line thing going up.

I: Okay. So I -- I won't -- you don't have to actually make [00:14:34] the line thing, I was just asking [Yeah] the sampler itself. So let's, uh, let's create that sampler.

S: OK. So then... and then, what should I put the repeat, as? Like, 10?

I: Don't worry about it -- don't worry about that.

S: Just, like, in general?

I: Yeah.

S: This one blue. And then color. Oops.

I: Okay. So, um, we're gonna do something a little bit new to kind of -- because we want to -- we're interested in what happens as we grow the sample, [mh] as we keep on adding [00:15:20] on. So normally TinkerPlots when you run it -- let me speed this up -- when you press run again you just get a totally different sample of 5. [mh] But what I want to do here is to keep on adding one on to whatever we have already. [OK] There is an option to do that. [OK] Um, so if you go into the options menu...

S: Right here?

I: Yep.

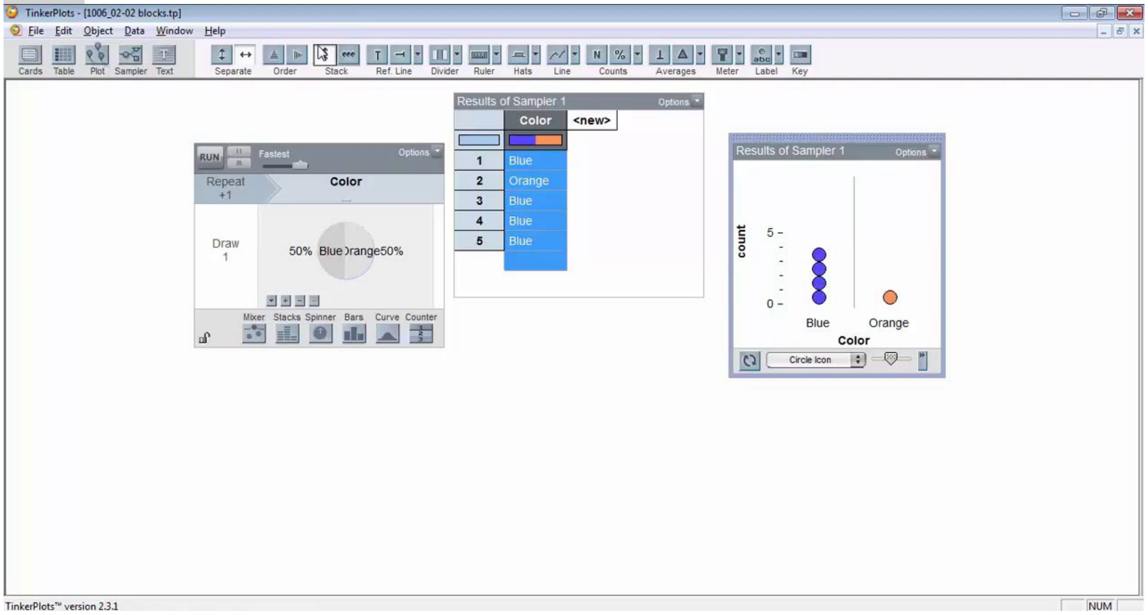
S: And then sampler options --

I: Sampler options. Um, we can turn *off* Replace Result Cases. [OK] So this means that [00:15:53] we're adding on [Oh, OK] to the sample each time. Um, so let's change this to 1. Um, let's plot the -- plot our sample so far.

S: Right now?

I: Yep.

S: OK.


TinkerPlots - [1006_02-02 blocks.tp]
File Edit Object Data Window Help
Cards Table Plot Sampler Text
Separate Order Stack Ref. Line Divider Ruler Hats Line Counts Averages Meter Label Key
RUN Fastest Options
Repeat +1
Color
Draw 1
50% Blue Orange50%
Mixer Stacks Spinner Bars Curve Counter
Results of Sampler 1 Options
count
5
0
Blue Orange
Color
Circle Icon
TinkerPlots™ version 2.3.1
NUM

Results of Sampler 1 Options

	Color	<new>
1	Blue	
2	Orange	
3	Blue	
4	Blue	
5	Blue	

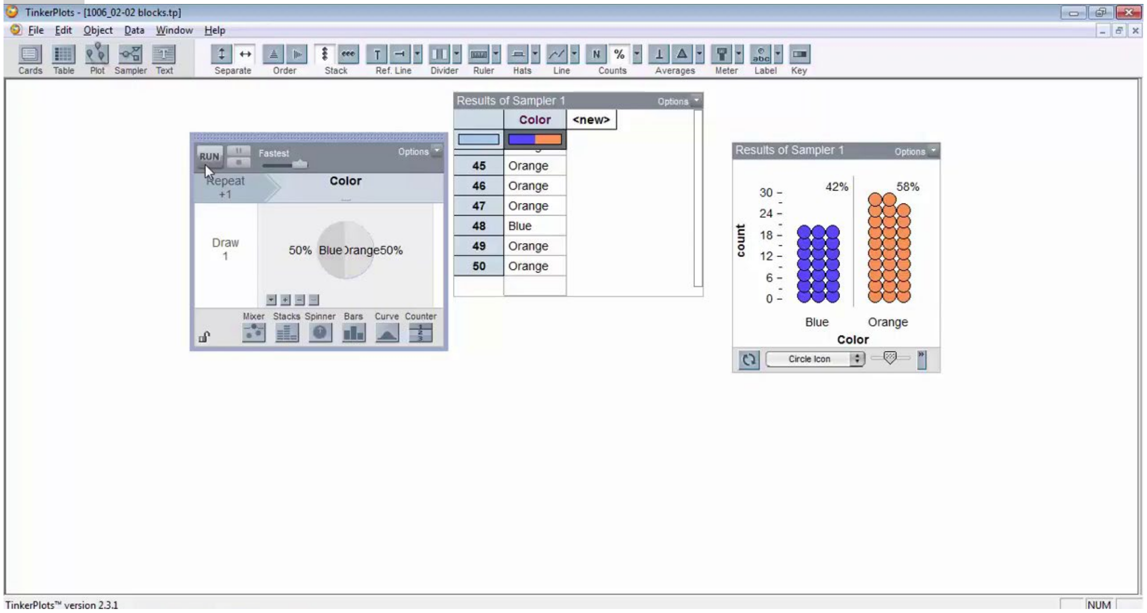
I: So we have -- oops. Um, so we can just stack those so we can see them. So you have blue and orange, um, and let's look at the perc-- the percentage blue. Um, so, just so we have kind of a clean start, just to start all the way from one, [mh] let's -- um, you -- let's actually delete these resul-- [00:16:34] these cases. So you can go to Options, and Delete All Result Cases. Yup. Just to give us a clean slate. [OK] And so now what I'd like to do is let's -- uh, let's actually go up to 50, um, just adding on each time. And, um -- yeah just take a look -- look at the percentage blue and sort of see how -- what's happening as we draw [OK] more and more samples.

S: So then I just press run?

I: Press run.

S: And then...

I: So we're at a [00:17:04] hundred percent now... so you can keep on clicking to see it grow. Okay. That's 50. Um, so, um, what did you notice so far, or can you describe what you saw happening with the proportion? [Yeah.]

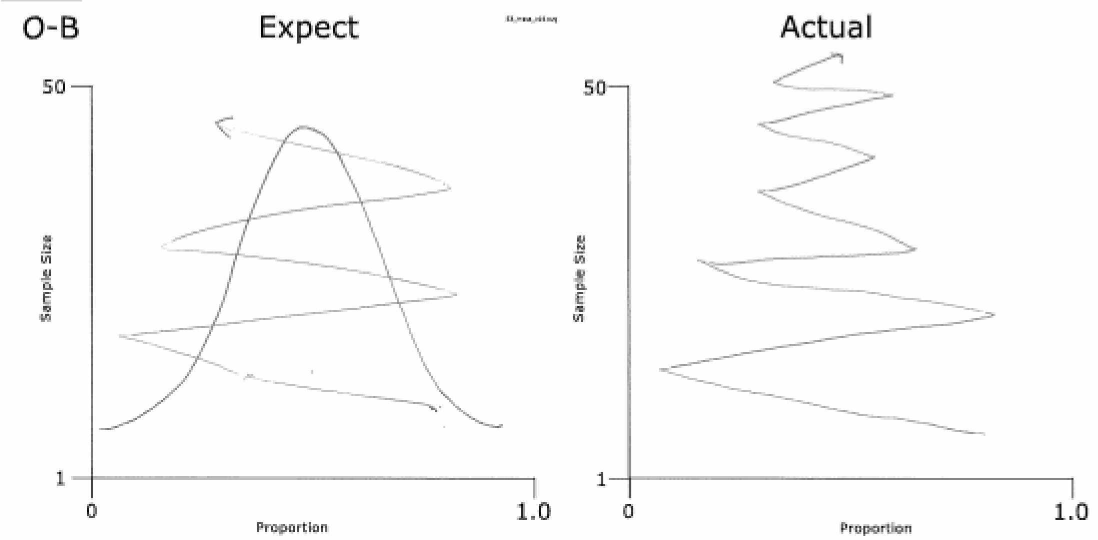

TinkerPlots - [1006_02-02 blocks.tp]
File Edit Object Data Window Help
Cards Table Plot Sampler Text Separate Order Stack Ref. Line Divider Ruler Hats Line Counts Averages Meter Label Key
RUN
Fastest
Options
Repeat +1
Color
Draw 1
50% Blue Orange 50%
Mixer Stacks Spinner Bars Curve Counter
Results of Sampler 1
Options
Results of Sampler 1
Options
count
30
24
18
12
6
0
42%
58%
Blue
Orange
Color
Circle Icon
TinkerPlots™ version 2.3.1
NUM

	Color	<new>
45	Orange	
46	Orange	
47	Orange	
48	Blue	
49	Orange	
50	Orange	

S: It was kind of going back and forth, like, it would get -- like, one side would get added, and then, like, if it, like, got added again, it would just keep on increasing or decreasing, so it was kind of going, like -- sometimes it [00:18:02] increased, sometimes it decreased, but, [mh] um, once we got like more times it started just decreasing, like, less. So instead of having, like, say, like, a 10% decrease or increase [Hm], like, it went more of, like, 2% or 5%, or something like that. I wasn't looking at the exact percentages, but. [OK] Um, then it kind of started going more towards, like -- instead of, like, say, like, 80%, it was more centered around, like, [00:18:26] 50/50. [OK] So.

I: Okay. Can you sketch what -- just about what happened, just now?

S: So then... I think it was on this side, we'll go with it.



I: OK. And was that -- was that what you expected to happen?

S: Yeah I didn't really think about the -- as mu-- as much as like in the -- the, like -- the expected one, um, that, like, it's a bigger -- like, I kind of talked about but, like, didn't really write it, um, how like it's a [00:19:01] bigger range in like -- w-- in like the zigzagging, and then it kind of, like, gets smaller as you go up, like, because there's less, like, back and forth.

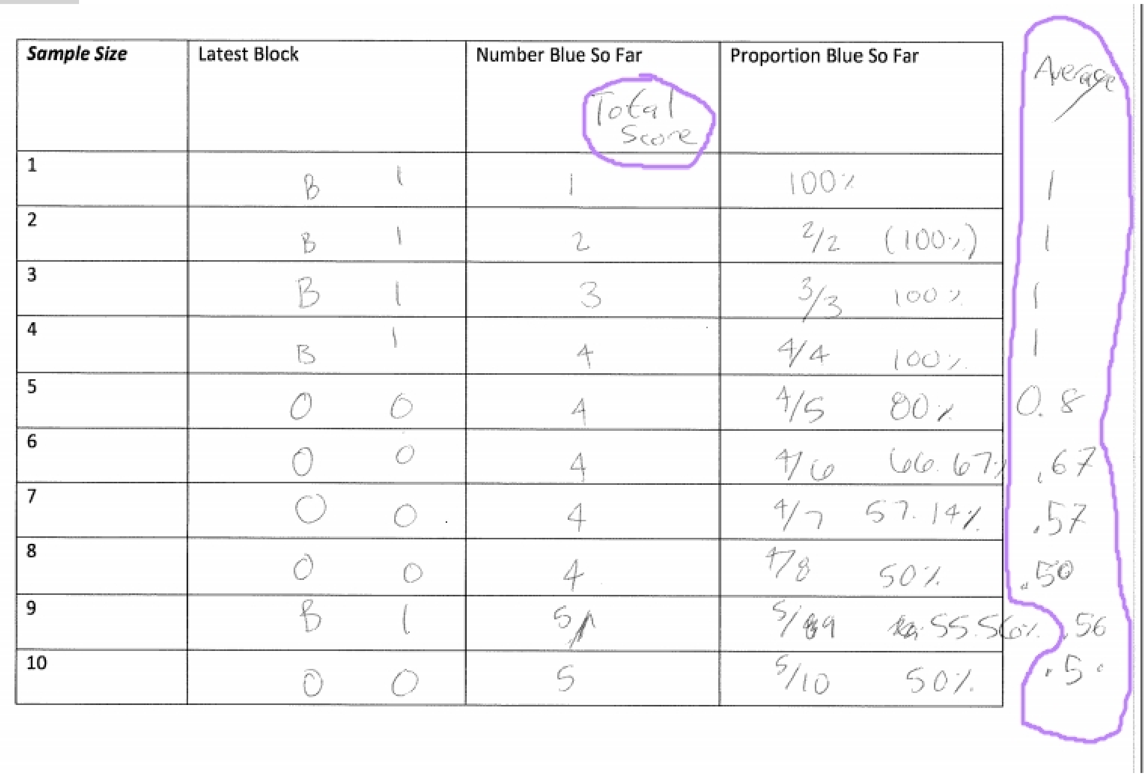
I: Okay. And so why does it get smaller as you go up?

S: Because the sample size increases so you're having, like, each time you put it -- one in, it'll de-- like, the proportion, like -- or the percentages won't change as much when you go in between like one and [00:19:26] two. [OK] So.

Growing Sample Means: 0-1

I: Okay. So we're gonna change one thing, um, um, to help us actually have TinkerPlots draw this for us. Um, so we're gonna model the same situation, so shaking the box and drawing one block, but now every time we draw blue we score it as a one, and every time we draw orange we score it as zero. So, um, I did this -- so I would score this as a zero, basically. Um, so how could you set up this zero one situation in [00:20:07] TinkerPlots?

S: Um, I think we could keep the sampler, and we would just put -- instead of having orange, we could have it as a one -- or zero, and then blue as one, um, and then that way, like, when you put it on, like, the plot, you'll be able to, like, look at, like, the percentages, or like proportions a lot easier, like, on the -- the line.



Sample Size	Latest Block	Number Blue So Far Total Score	Proportion Blue So Far	Average
1	B 1	1	100%	1
2	B 1	2	2/2 (100%)	1
3	B 1	3	3/3 100%	1
4	B 1	4	4/4 100%	1
5	0 0	4	4/5 80%	0.8
6	0 0	4	4/6 66.67%	.67
7	0 0	4	4/7 57.14%	.57
8	0 0	4	4/8 50%	.50
9	B 1	5	5/~~8~~9 ~~ka~~ 55.56%	.56
10	0 0	5	5/10 50%	.5

I: Okay. Um, so now we're gonna be interested in the mean [00:20:36] instead of the proportion blue as the sample size increases. Um, so -- um, so if we went back on this sheet, and we were gonna kind of do it the new way looking at the mean. So, um, we still have the number blue, and so this is just the total so far, kind of our total score. So we got a blue -- and these are ones, and these are zeros, which look a lot like Os, too /cs/. And our total score so far, we got one, and then our

[00:21:15] total score we got up to 2, to 3, then to 4, and then it stayed at four /c/, um, through here, and then it bumped up to 5 and stayed at 5. And then our average is -- um, just going to be 1 out of 1, here, so it'll just stay at 1, through these 4. Um, here I have, um, four out of five so that's point eight, point six seven, because it just stayed there, we're kind of [00:21:46] going down a little bit. Point five seven, and then my total score of eight, divided by my total number of scores is that, so again at fifty percent -- or 0.5. And then point five six, and point five. Um, so just because through the rest of these activities, we're going to be really talking about, um, talking about averages, so I want to do this in terms of averages. Um, so can you sketch what you think the -- now [00:22:24] going up to 200. Can you sketch what you think the sample size plot would look like for the mean of this kind of zero one situation.

S: Yeah. It'll probably... let's see. Um... So it's the mean. I guess.. does it really matter what side I start from? =Or, just like -- it's more like --=

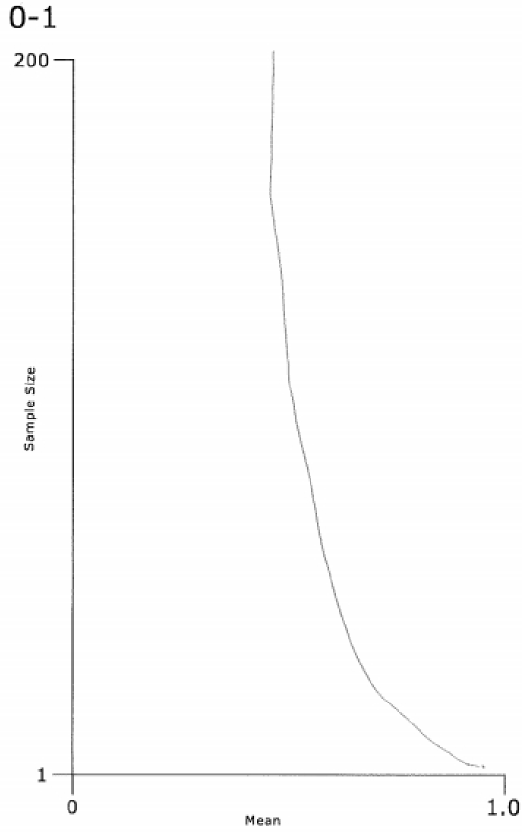
I: =You -- you can do whatever --= it's just [OK] what might happen one time. Yeah.

S: Okay. Um, so, like, if it [00:23:07] happens just, like, one time it'll probably go up, and it's kind of like -- stay at like fit-- like point five [OK], I would say. =Like, as you do more.=

I: =And why 0.5?=

S: Um, because that's like 50/50 per -- like, 50% chance, like, you're gonna get orange, or you're gonna get blue.

I: Okay. And can you just describe kind of what you drew, =a little bit more?=



S: =Yeah, I started= at one side and I just kind of did, like, a curve to like towards -- that would be, like, the end of it would [00:23:40] be at 0.5 [OK] on the X axis.

I: Okay. Um, so, um, so let's, uh -- let's actually create this in TinkerPlots.

S: Okay. So that -- oh, sorry /yawns/ -- would this just be one, then for --

I: Sure. [OK] You can make that one and then zero. We can delete those old --

S: So then --

I: Or just -- maybe just delete the whole table [OK] just cuz it's gonna be a little confusing. [OK] OK [Yeah, so..], so let's just do from scratch.

S: And then just keep =doing it?=

I: =Yeah, so= you can do a couple more times -- we just need to wait -- okay so that's [00:24:22] good.

S: /inaudible/

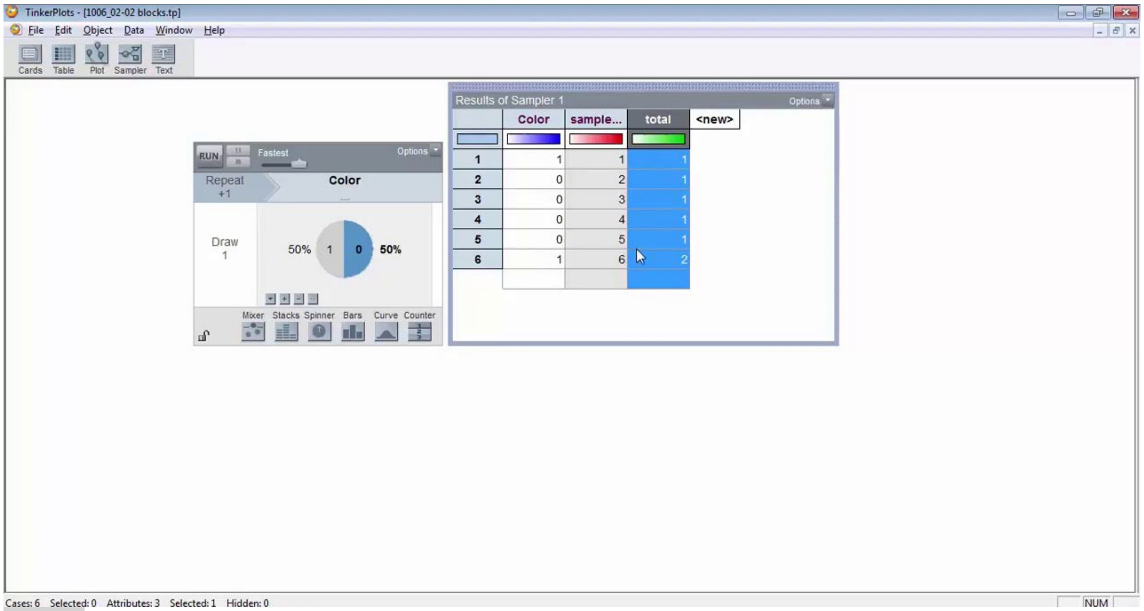
I: So we have some examples, there. [Yeah] Um, so we're gonna add some formulas so we can actually create this plot in TinkerPlots. Um, so, um, first we're gonna add a column just to get the sample size. Um, so you just type -- call the column Sample Size. Um, and then we're going to -- so you can click Edit Formula, and there's a function called caseindex --

S: =Which /inaudible/...=

I: =Which you can= just type. You can just type it. caseindex.

S: So just one [00:24:54] word?

2.621:
2.622: I: Yep. And basically that just gives
2.623: you whatever kind of row you're on. So it --
2.624:
2.625: S: OK,
2.626: then just do =okay?=
2.627:
2.628: I: =OK.= And this is nothing fancy,
2.629: it just gives you [mh] exactly that. [OK] But
2.630: we can't drag that down onto a plot [OK], so we
2.631: sort of create this new column. Um, so the
2.632: next one I'll do. We want to kind of [OK]
2.633: recreate this kind of total column. [OK] So
2.634: we'll get the total score so far.
2.635: Um, the way I'm doing it is basically --
2.636: [00:25:25] I'm just gonna incrementally add on
2.637: whatever we got. [OK] So the previous
2.638: total, whatever that is, and then add on --
2.639: you've called this Color so that's the
2.640: result of our current, um, time. So we just
2.641: keep on adding on [OK] whatever we got the
2.642: latest time. Um, and this stays at one.. I can /I clicks until another 1 appears/ -- yeah.
2.643: Just to [OK] give this example. Um, so the
2.644: total so far is, um, we got one at the
2.645: beginning then it stays at one and then
2.646: [00:25:59] finally we get another one. And [OK] the to--
2.647: total goes up to two. Um, so what would the average
2.648: be -- so we'll just kind of talk through this for
2.649: a moment. What would the average be here
2.650: at one?
2.651:


TinkerPlots - [1006_02-02 blocks.tp]
File Edit Object Data Window Help
Cards Table Plot Sampler Text
RUN
Fastest
Options
Repeat +1
Color
Draw 1
50%
1
0
50%
Mixer Stacks Spinner Bars Curve Counter
NUM

Results of Sampler 1

	Color	sample...	total	<new>
1	1	1	1	
2	0	2	1	
3	0	3	1	
4	0	4	1	
5	0	5	1	
6	1	6	2	

Cases: 6 Selected: 0 Attributes: 3 Selected: 1 Hidden: 0

S: Um, one... right. Yeah, one.

I: It's just one out of one, yeah. And then what would the average be at four?

S: Um, point two five.

I: Okay. And then here at six?

S: And then two out of six -- 1/3, so 33 percent.

I: Okay, so, um --

S: Or 0.33, I guess.

I: So how are you -- how do you calculate -- for [00:26:38] each row, how do you kind of calculate the mean?

S: Um, I did the total divided by the

sample size.

I: Okay. So let's go -- add a new column for the mean [OK] that just has that same calculation.

S: So then it would just be the total divided by sample size --

I: Double-click. There you go. Okay. Um, so let's -- um, now let's recreate that same plot -- the sort of, the -- the sample size plot that we did before.

S: OK, so then -- should I put the sample [00:27:25] size on the Y axis?

I: Yup. Yup, just like before.

S: And then what do we want on the -- n--

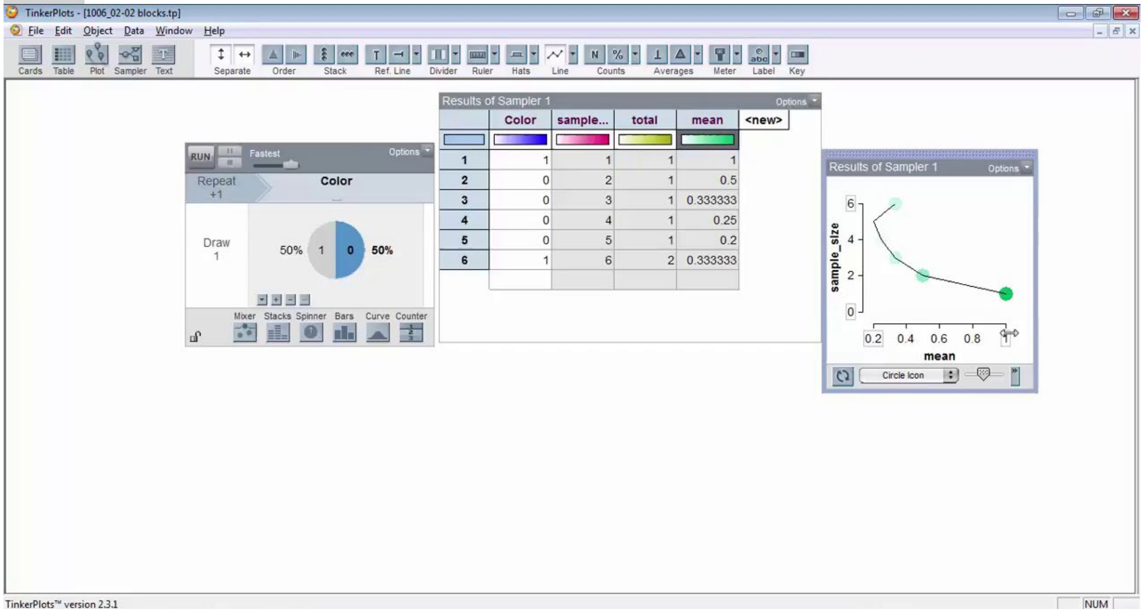
I: So the mean, now --

S: The mean? OK. And then should I separate them?

I: Yeah. Drag -- you'll have to drag that -- /c/ shoot. You'll have to drag it up. /c/

S: OK.

I: And then we can do the -- change the icon to Borderless. And then add the line [OK], to sort of see things as they grow. OK. Um, so, um, now let's actually grow it up to -- up to 200. So you can click on run to kind of [00:28:06] watch it grow.


TinkerPlots - [1006_02-02 blocks.tp]
File Edit Object Data Window Help
Cards Table Plot Sampler Text
Separate Order Stack Ref. Line Divider Ruler Hats Line Counts Averages Meter Label Key
RUN
Fastest
Options
Repeat +1
Draw 1
Color
50%
1
0
50%
Mixer Stacks Spinner Bars Curve Counter
Options
sample_size
6
4
2
0
0.2 0.4 0.6 0.8 1
mean
Circle Icon
TinkerPlots™ version 2.3.1
NUM

Results of Sampler 1

	Color	sample...	total	mean	<new>
1	1	1	1	1	
2	0	2	1	0.5	
3	0	3	1	0.333333	
4	0	4	1	0.25	
5	0	5	1	0.2	
6	1	6	2	0.333333	

Results of Sampler 1

S: So I -- should I just go --

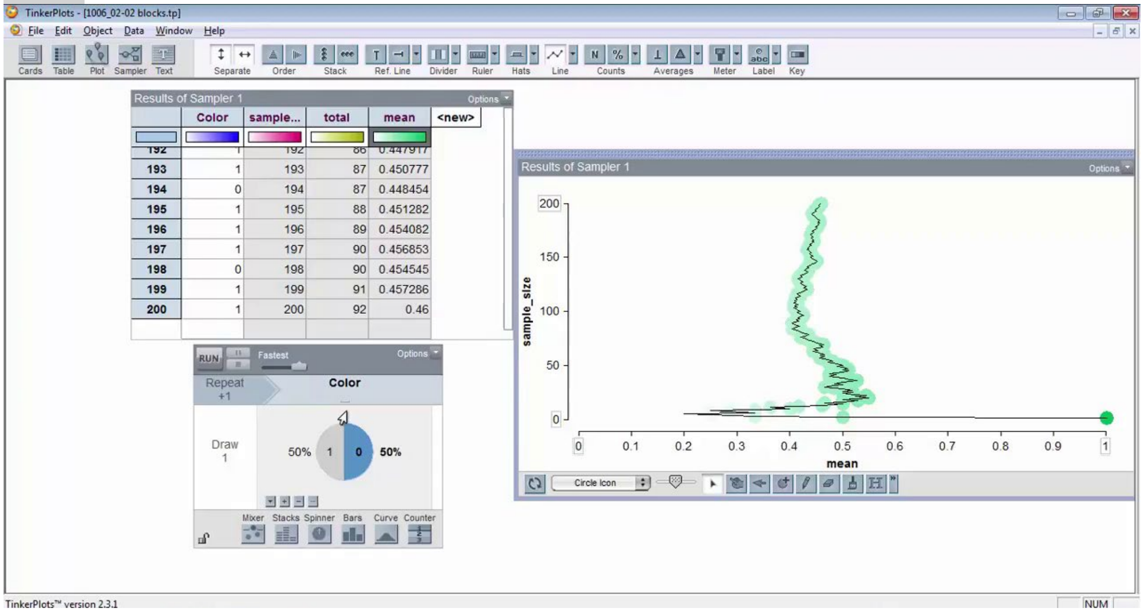
I: Keep clicking. [OK] Yep. [/c/] So you can just watch the plot as it grows. And actually before we do that, let's just change this to 0 --

S: 0? OK.

I: -- so we see the full range. [OK] Okay. Um, so what do you notice?

S: Well after it was like -- it was like at -- w-- you-- it's gonna be at either one or zero in the beginning [mh], and then it's kind of gonna -- maybe to -- maybe shoot over at one point, like, /c/ [00:29:09] and then it's gonna stay around, um, I'd say, like, it'll kind of be like around 50 after about, like, it seems like at least on this graph, like, it's gonna be like around like 25 or 30 [OK], and then it's kinda gonna just kind of stay in

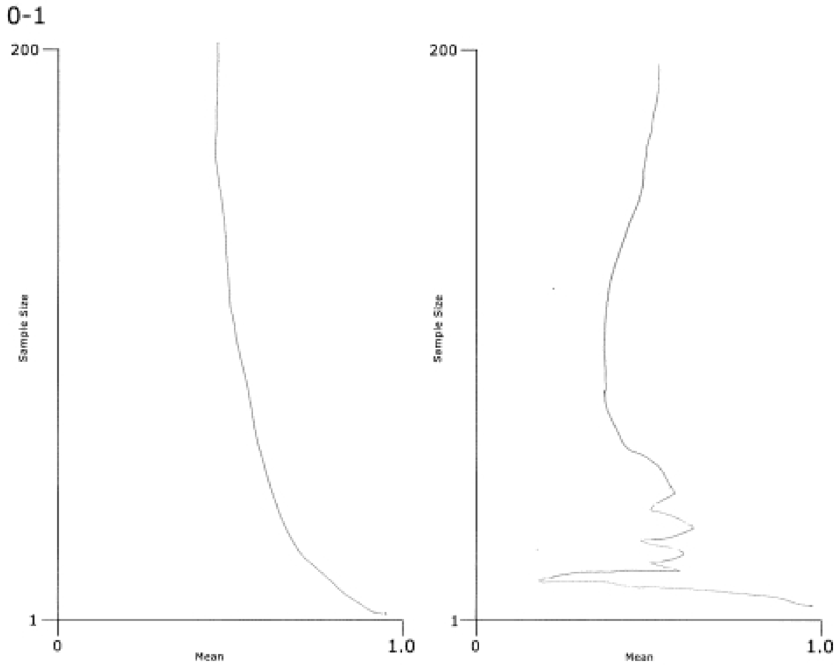
2.742: the middle. [OK] So.
2.743:
2.744: I: We can make that a little
2.745: bit bigger, let's just make that bigger. [OK] Just so we can see it. Um, so, um, so you can sketch the plot
2.746: that you actually saw. [OK]
2.747:
2.748:


TinkerPlots - [1006_02-02 blocks.tp]
File Edit Object Data Window Help
Cards Table Plot Sampler Text Separate Order Stack Ref. Line Divider Ruler Hats Line Counts Averages Meter Label Key
Results of Sampler 1
Options
RUN
Fastest
Repeat +1
Color
Draw 1
50% 1 0 50%
Mixer Stacks Spinner Bars Curve Counter
sample_size
200
150
100
50
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
mean
Circle Icon
TinkerPlots™ version 2.3.1
NUM

Results of Sampler 1

	Color	sample...	total	mean	<new>
192	1	192	86	0.447917	
193	1	193	87	0.450777	
194	0	194	87	0.448454	
195	1	195	88	0.451282	
196	1	196	89	0.454082	
197	1	197	90	0.456853	
198	0	198	90	0.454545	
199	1	199	91	0.457286	
200	1	200	92	0.46	

S: { /inaudible/ }



I: [00:29:55] Um, and does that match what you expected?

S: Somewhat. I mean, it's, like, slightly different, cuz -- I didn't really think about, like, how it's, like, definitely gonna either start at one or zero, can't really start, like, in the middle. [mh] So, just how it's gonna start. And then probably zigzag in the beginning [mh], but it makes sense. [OK] I just didn't, like, think about the -- I did more of, like, well it's about [00:30:17] gonna be like this, and not, like, really zig zag. So. [mh]

I: Um, what would happen if we started over and did this again?

S: Um, it'll probably stay around the same. Like, it's probably gonna still, like, be a lot more zig zaggy at the bottom than it is like at the top, where it's kind of gonna just kinda taper [OK] off into a straighter line.

I: And so why is that? Why is it more zig zaggy at the [00:30:41] bottom?

S: Because there's small-- a smaller sample size. So each time you put, like -- have one more trial, um, it's gonna have, like, one less -- or I mean it's gonna have a bigger impact on the total, like, percentage or, like, say on -- in this case, like the mean. Um, it's just gonna be able to vary a lot more, while if you have like 200, then, like, adding one blue or one orange is gonna make less of a difference in, like, [00:31:10] say the percentages of each one, so. [OK]

Growing Sample Means: 0-1-1

I: Um, so now suppose, um -- we -- there's one orange and two blue blocks -- shoot -- um -- {nope, that's oran-- /cs/}. Two -- one orange and two blue blocks in the box. So again each blue block is scored one point, and each orange block is scored as zero points. Um, what would you expect to see in terms of the mean score as you keep drawing blocks?

S: I'd say instead of, um, being centered around, [00:31:50] like, 50, it's gonna be probably a little bit -- it's gonna be more close to, um, 1 [OK] or 0.5. It's gonna be more close to, like, one say, than in the s-- in -- in -- in the middle, so.

I: Okay. So what would you expect to be -- I think you were already starting to answer this, but would you expect to be different from what you saw already?

S: Um, so it'll probably be -- instead of, like, being centered around this way, it's gonna be more of, like, [00:32:23] centered on, like, the right side of the graph, and I think the zigzag will probably be, like, the same. Like, it's still gonna vary a lot in -- at the very bottom, but it's gonna be centered slightly to the right.

I: Okay. Um, so how can you change the model on the TinkerPlots?

S: Um, should I just -- like, could I put, like, 75 and 25%, is that the right percentage, =or -- or --=

I: =So there's= two blue blocks and one orange block. [00:32:53]

S: I guess we could do the proportion. Would that be betterrrr? Or... my brain is a little fuzzy right now.

I: That's okay. That's okay. So, um --

S: It'd be like -- s---

I: What's your chance here of drawing an orange block?

S: 33%.

I: Okay.

S: So then. [OK] All right. Or should I -- yeah, is that fine, even though it's, like, because it was 33 -- it would be, like 33.33 [=If--=], is it fine?

I: If you're -- if you think it's close enough, =I think it's close enough.=

S: =I think it's close enough.= /c/

I: Yup, I'm -- I'm fine with that.

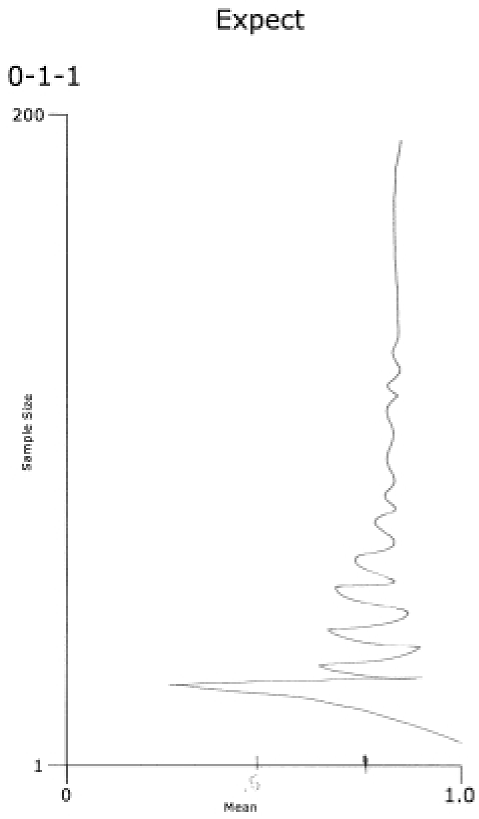
S: Okay.

I: Um. Okay, so let's see what that [00:33:37] looks like. So first, let's, yeah -- let's [OK] get rid of the old ones.

S: {/inaudible/, OK}

I: Oh. And I'd like you to sketch, before we actually do it, what do you think you might see in terms of [OK] this one.

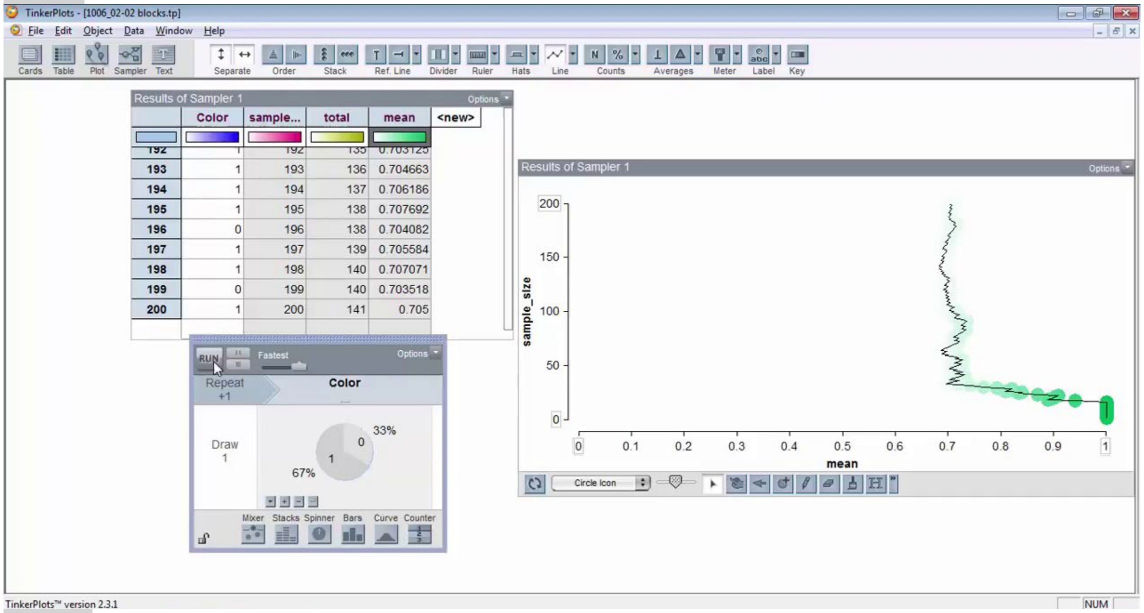
S: So we'll say, like, that's like, the /inaudible/. It's, like, point 5. So it'll probably be like... sh.... or, yeah, close enough. It's closer to one. Maybe -- not like exactly one, but /c/ yeah.



I: Okay. OK. Is there something -- do you have some [00:34:10] hesitation about this? Or you were saying, um, e-- I'm just reacting because you -- you -- it seemed like you had some hesitation about the plot that you just drew, or...

S: Oh, just like, I did a little bit t-- more to the right. I was gonna try to make it more like centered like [Oh, OK] at, like, right there. [OK, gotcha, gotcha.] I'm not great at drawing. /c/

I: That's all right. I -- I'm terrible. /c/ Um, okay so now let's see what that actually looks like. [OK] So what do you notice now, looking [00:35:18] at this plot?


TinkerPlots - [1006_02-02 blocks.tp]
File Edit Object Data Window Help
Cards Table Plot Sampler Text Separate Order Stack Ref. Line Divider Ruler Hats Line Counts Averages Meter Label Key
Results of Sampler 1
Options
RUN
Fastest
Repeat +1
Color
Draw 1
33%
0
1
67%
Mixer Stacks Spinner Bars Curve Counter
Results of Sampler 1
sample_size
200
150
100
50
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
mean
Circle Icon
TinkerPlots™ version 2.3.1
NUM

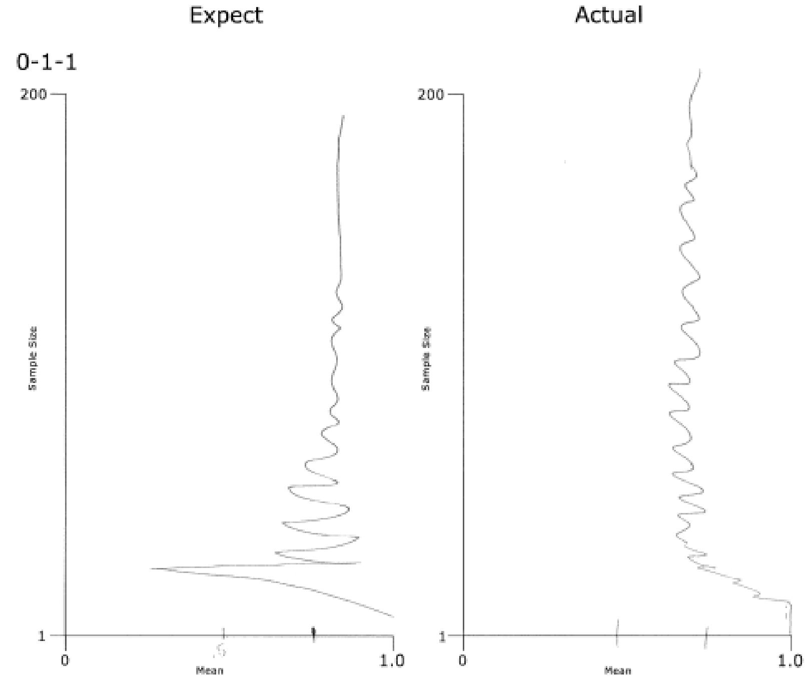
Results of Sampler 1

	Color	sample...	total	mean	<new>
192	1	192	135	0.703125	
193	1	193	136	0.704663	
194	1	194	137	0.706186	
195	1	195	138	0.707692	
196	0	196	138	0.704082	
197	1	197	139	0.705584	
198	1	198	140	0.707071	
199	0	199	140	0.703518	
200	1	200	141	0.705	

S: Um, it didn't really zigzag any farther than, like, 0.7. [OK] Um, I think partially it was because of how many, um -- we got so many blue blocks right in the very beginning, [mh] so beca-- it would -- it like -- it's going to va-- like, there's really -- like for this one, for the one, um, before with only two blocks -- um, the next one was orange, so, like, it would -- it'd like -- drastically changed, [Hm] while here, it's like, less of [00:35:48] a sample size, like, once I did get a blue-- once we did get a blue block, it didn't change quite as much. [OK] So there wasn't really as much as zigzagging, but it definitely pretty quickly it went, um, to the point s-- like, around point sevenish.

I: Okay. Okay. So can you sketch what you actually saw?

S: Yeah.



I: And, um, what was -- um, so first, like, what was similar to what [00:36:28] you saw when you had the -- had the one orange and one blue block, or -- [mh] when we had the zero and the one here.

S: Um, I would say probably, like, the first, like, the two -- first, like, couple of trials, like, it definitely goes back and forth a lot more. Um, but it get -- like, has a smaller, like, difference at the very top.

I: Okay. And what's different about the two graphs?

S: Um, that one is more centered around 0.5, [00:36:55] while, this one's more of, like, point seven, point seven five. [OK] So.

I: Um, and, uh, what, uh -- what do you think would happen -- what do you think might happen if we started over and did this again?

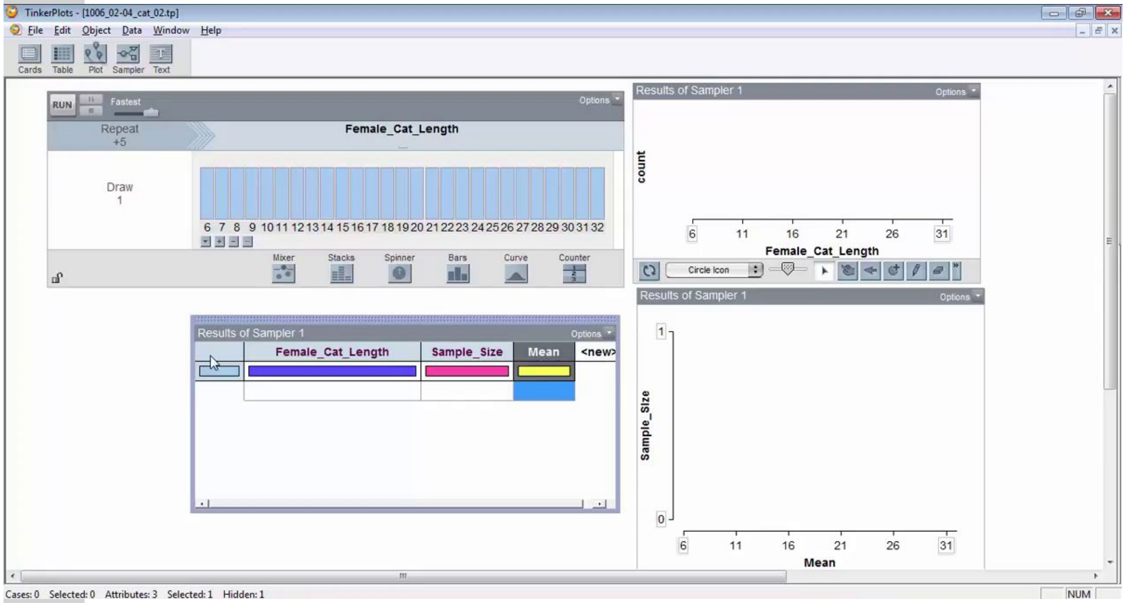
S: Um, I think depend-- if we -- like a couple of orange blocks in the beginning, um, I'd say it'd be more zigzaggy most likely I would say, but there's also like h-- for this one, it's like a higher percentage of blues [00:37:21] compared to the oranges so it makes sense, how they're like -- it's gonna be a little bit more towards like the blue side, so.

Growing Sample Means: Cat Factory 1

I: Okay. Um, so do you remember an activity from class where you created a Cat Factory?

S: Yeah. I remember using the -- from -- like the little s-- sheet or, like, go to TinkerPlots thing.

I: Um, so we're gonna create a-- um, we're gonna, um, create another Cat Factory now where you can [00:37:50] create female cats of different lengths. It's a little bit less complex that what you were doing [Yeah] in class. Um, so here you'll see that the cats have, um, different possible lengths between 6 and 32 inches. Um, the bars represent how likely each length of cat will be. So -- that here all cats are equally likely. Um, how reasonable does this seem as a model for creating cats? [00:38:18]


TinkerPlots - [1006_02-04_cat_02.tp]
File Edit Object Data Window Help
Cards Table Plot Sampler Text
RUN Fastest Options
Repeat +5
Female_Cat_Length
Draw 1
6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
Mixer Stacks Spinner Bars Curve Counter
Results of Sampler 1 Options
count
6 11 16 21 26 31
Female_Cat_Length
Circle Icon
Results of Sampler 1 Options
Female_Cat_Length Sample_Size Mean <new>
Results of Sampler 1 Options
1
Sample_Size
0
6 11 16 21 26 31
Mean
Cases: 0 Selected: 0 Attributes: 3 Selected: 1 Hidden: 1
NUM

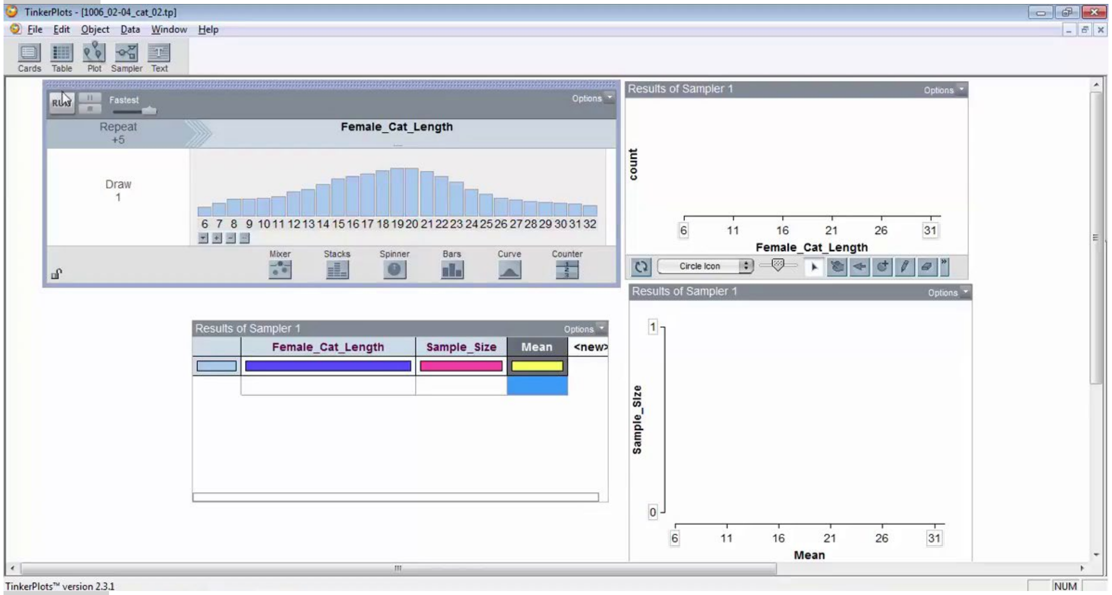
S: I mean -- it's, like, random, so it makes sense, but most likely whatever, like, the -- like -- for, like, say, like, a certain, like, breed or something, like, they might tend to be longer, so it kind of also, like, depends, like, cats are probably more likely to be, like, say, like, 22 rather than, like, 6 inches or whatever -- or 6 feet -- or no not-- definitely not, it'd be =six -- six inches.=

I: =That's a scary cat.= /cs/

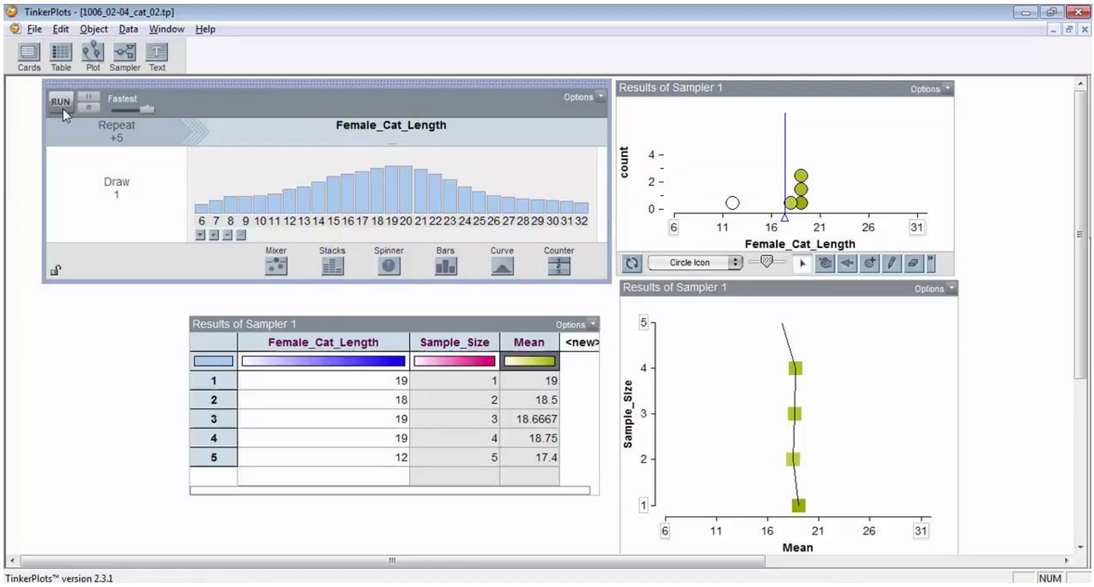
S: Yah. A tiger. /cs/ Um, but it's probably gonna be, like, centered more [00:38:48] around, like, what most cats are gonna be, so it's probably not gonna be as many, like, on each end, it'd probably be more like a bell curve, like, centered one way or the other. So.

I: Okay. Um, so you can draw on the sampler window whatever likelihood seems reasonable. So if you bring the mouse to the

2.1012: sampler window [Yeah] you can see that you can
2.1013: just draw, click and draw.
2.1014:
2.1015: S: We'll go-- oh -- we'll go like that. /c/
2.1016:



2.1017: I: OK. Um, so you can
2.1018: the Run button [OK] once. And so the s-- the
2.1019: [00:39:22] file's set up so that the mean appears
2.1020: here, it's already set up right next to
2.1021: the cat length. And you can see a plot of the
2.1022: cat length, um, in the sample here. So this is
2.1023: just the plot [mh] of the individual sample.
2.1024: And then here, um, is the plot of the means
2.1025: by the sample size, kind of what we were
2.1026: drawing before. [OK] Um, so this square shows
2.1027: that the mean at sample size two was
2.1028: eighteen point five. [OK] Um, so what do you
2.1029: [00:39:52] think the graph would look like if you
2.1030: kept growing the sample?
2.1031:



Results of Sampler 1

	Female_Cat_Length	Sample_Size	Mean	<new>
1	19	1	19	
2	18	2	18.5	
3	19	3	18.6667	
4	19	4	18.75	
5	12	5	17.4	

S: Um, I guess I think at the very end it'll probably be, like, around like the 20ish area, kind of like where it started, just because, like, it's more likely to have, like, a cat that's, like, the average -- like, say 20 or something like that, than it is to have, like, a cat that's like 32 inches or something. But, um, I think it'll probably [00:40:19] kind of like stay in that area, maybe like zigzag around, maybe. But like not a huge difference, I don't think.

I: Okay. So, um, so there's the -- um... so I'm sorry I just realized I had a question.

S: No, that's OK.

I: So say what you -- say again what you said about not a huge difference or not?

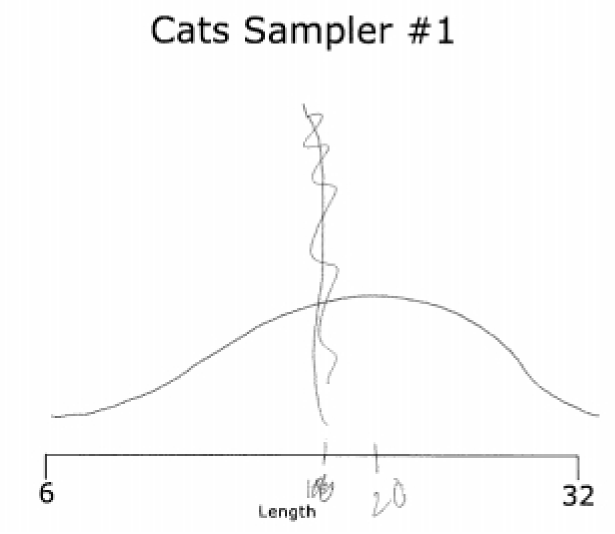
S: Well, in, like, say, like, the 1 to 2 -- or like

1 to 0, it can't really be anything in between, it can only be one thing or [00:40:55] another. [Hm] While this one it can be, like, multiple different things, [mh] so it won't be zigzagging quite as much, like, all around, it'll be more of, like, um, kinda staying in the middle. And also it's taking the average -- or it -- like the -- the me-- it's, like the mean, so it's not going to be, like, varying too much, because it's taking that mean of everything, so.

I: Okay. So can you draw what you actually have already in your sampler [OK] [00:41:26] up on top here.

S: So then... we'll say it was like 18, so I was, like...

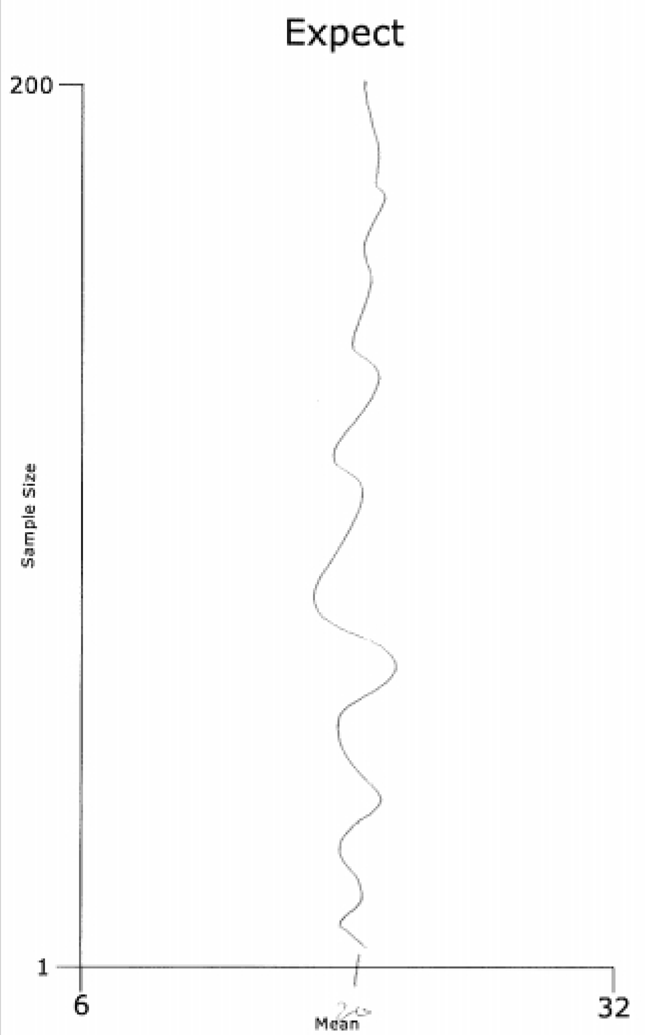
I: Um, so I'm sorry. [OK] To draw actually this [OK] the sampler there. /c/



S: {That's OK.} Close, enough, we'll say.

I: Okay. [OK] Um, and now what do you think the graph would look like if you kept

2.1083: growing the sample from one [OK] all the
2.1084: way to 200.

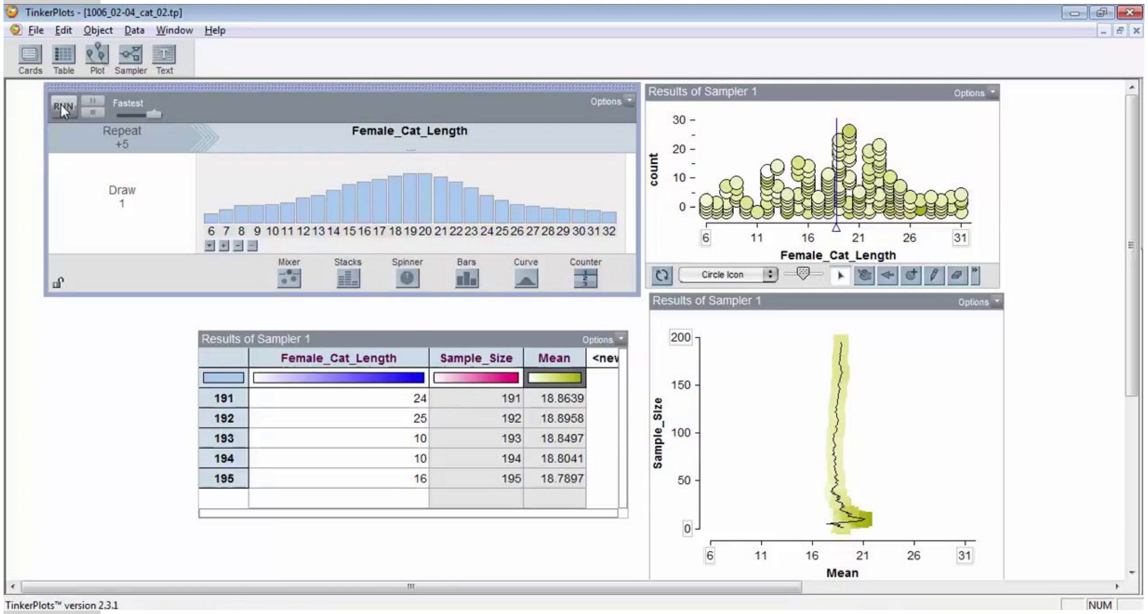


2.1085:
2.1086: S: Um, I kind of just put this as,
2.1087: like, t-- 20-ish, because that's kind of where
2.1088: I have the highest point there. [OK] I
2.1089: think it'll kind of, like, just kind of
2.1090: [00:42:05] go back and forth, like, not just as-- it's kind of
2.1091: like went a little crazy right there, but
2.1092: I think it'll probably stay pretty close
2.1093: to, like, what the -- I put like the average
2.1094: as, so.
2.1095:

I: Okay. Um, so let's see what actually happens.

S: = /inaudible/ =

I: = So it's going by fives=, now so it will [OK] go a little faster. So what did you notice?

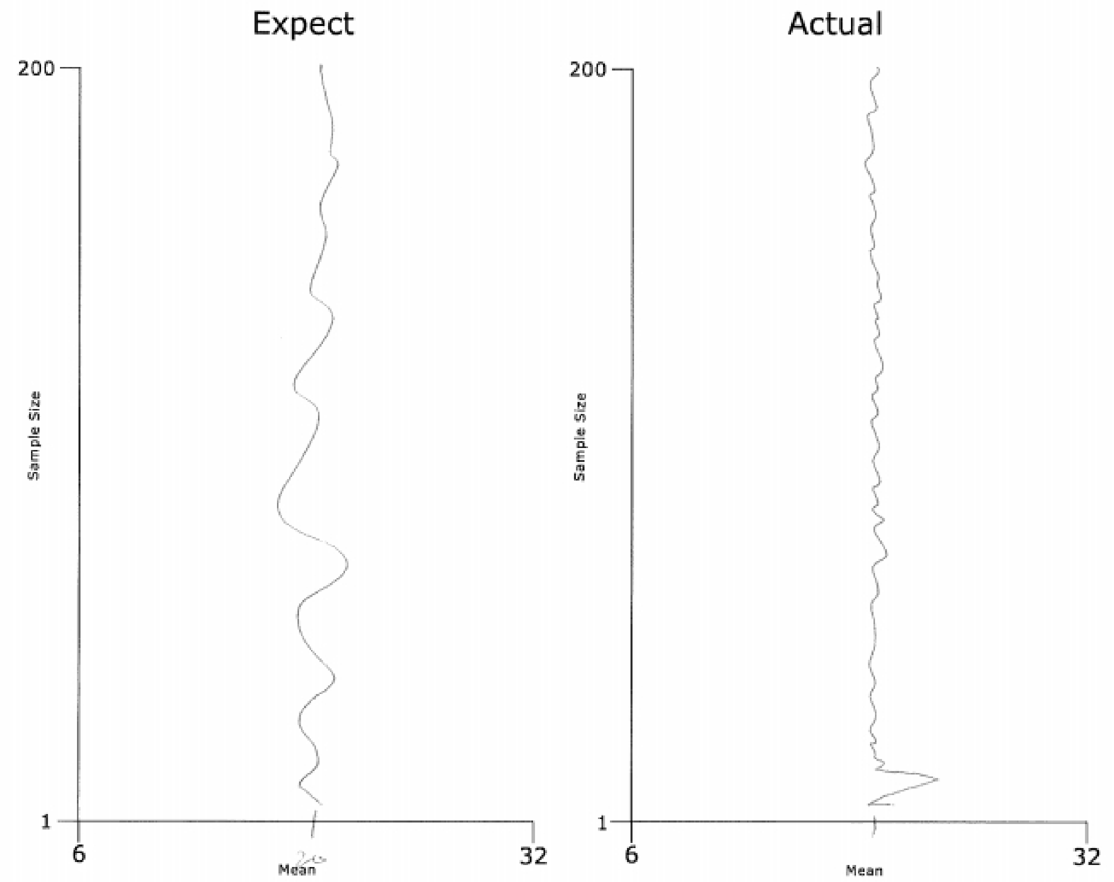


Results of Sampler 1

	Female_Cat_Length	Sample_Size	Mean	<new
191	24	191	18.8639	
192	25	192	18.8958	
193	10	193	18.8497	
194	10	194	18.8041	
195	16	195	18.7897	

S: There's a slight zigzag at the bottom, but it basically stayed, like, almost like a straight line [OK] [00:42:45] the entire time. So.

I: OK. Um, so sketch what actually happened. And what do you think would happen if we kept going?



S: It'd probably be -- get pretty close to like a straight line, even less like of a little zigzag, so.

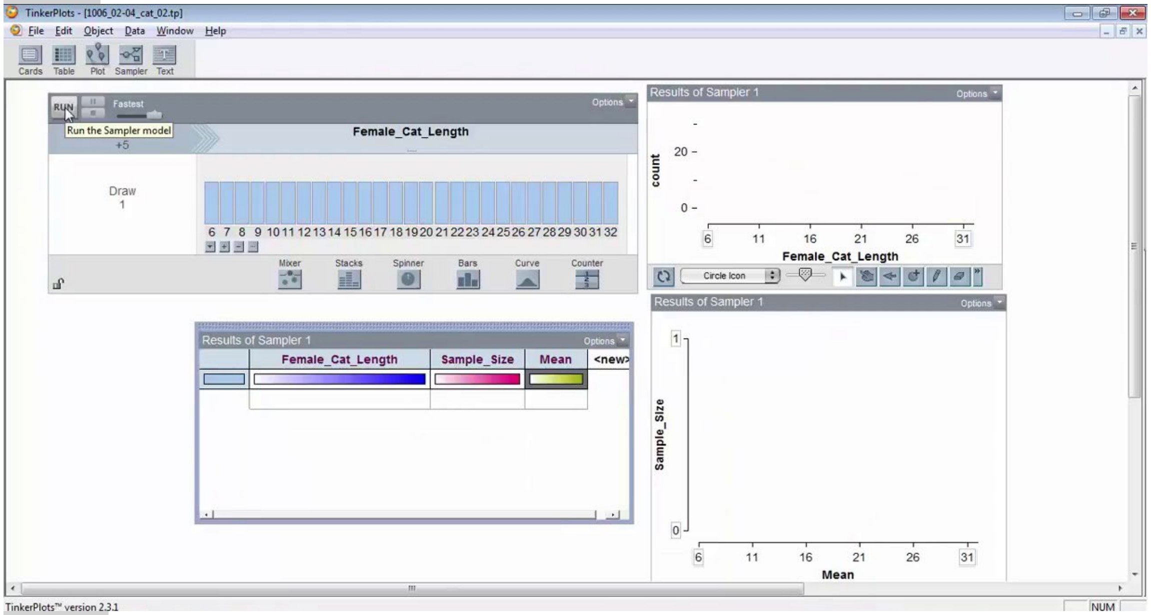
I: Okay. And, uh, what would happen if we started over and did it again?

S: Um, I think it would be about the same. There might be, like, a slight, like, zigzag at the very bottom just because [00:43:21] of the smaller sample size, kind of how it was here. Um, it might happen, it might not happen, but it'll probably still be, like, centered kind of around the 20 area.

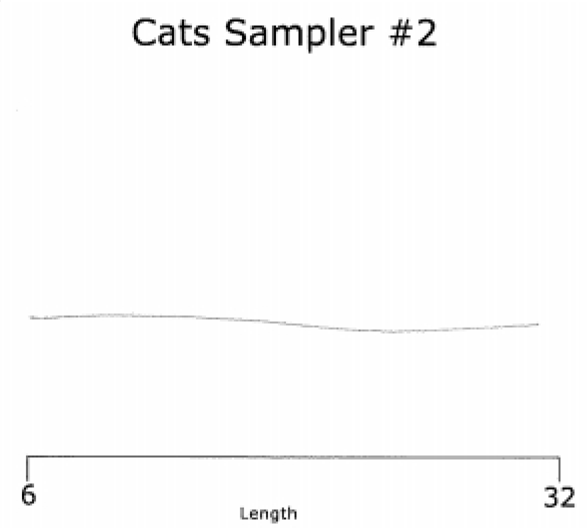
2.1132: Growing Sample Means: Cat Factory 2

2.1133: I: Okay.
2.1134: So now I'm gonna ask what can you -- what
2.1135: could you change in, um, the sampler -- um, in terms
2.1136: of the heights of the bars,
2.1137: what could you change to make the graph
2.1138: come out differently?
2.1139:
2.1140: S: Um, I could make it,
2.1141: like, more even I guess, cuz that way, like,
2.1142: [00:43:56] it'll be less, like, likely to have to the --
2.1143: like, the mean centered around, like,
2.1144: 20 or 19 or whatever it is. [OK] But, yeah,
2.1145: I probably say just like making them
2.1146: more even, maybe? [OK] Like,
2.1147: at least for the height of, like, the -- the ba--
2.1148: each bar.
2.1149:
2.1150:
2.1151: I: Okay. And how would that change the
2.1152: sample size plot here?
2.1153:
2.1154: S: I think it would
2.1155: zig zag more, um, just because there's like
2.1156: [00:44:30] -- a more likely chance that, like, each o-- e--
2.1157: like, different lengths are gonna be
2.1158: drawn. [OK] Rather than, like, right now
2.1159: it's saying that, like, 20 is, like -- gonna
2.1160: most likely happen, while if they
2.1161: were more even, like, it'll be more likely,
2.1162: like, say that, like, you could have the 30
2.1163: inch cat [Hm]. Um, so.
2.1164:
2.1165: I: OK. Um, so let's actually do that.
2.1166: We can -- so you can draw it evenly.
2.1167:
2.1168:
2.1169: S: Even enough. /cs/ [OK] It'll deviate a little
2.1170: [00:45:10] bit, so.
2.1171:

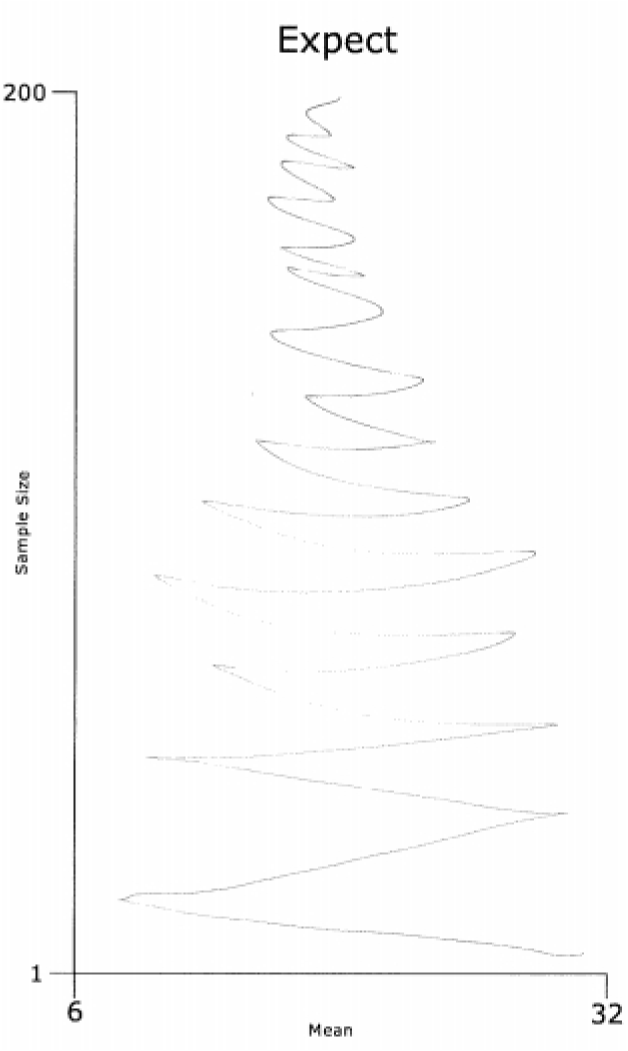
2.1172: I: We can, uh -- we can actually make it -- do
2.1173: you want to make it exactly even?
2.1174:
2.1175: S: Sure. /c/
2.1176: Why not.
2.1177:
2.1178:
2.1179: I: Okay. So now that's exactly even again. Um,
2.1180: so we can --
2.1181:
2.1182: S: Delete Case Results...
2.1183:
2.1184: I:
2.1185: Delete the case results. Um, yeah, can you draw your
2.1186: sampler?
2.1187:



2.1188: S: { /inaudible/ } The line.

2.1189: 

2.1190: I: And then what you'd
2.1191: expect to see.



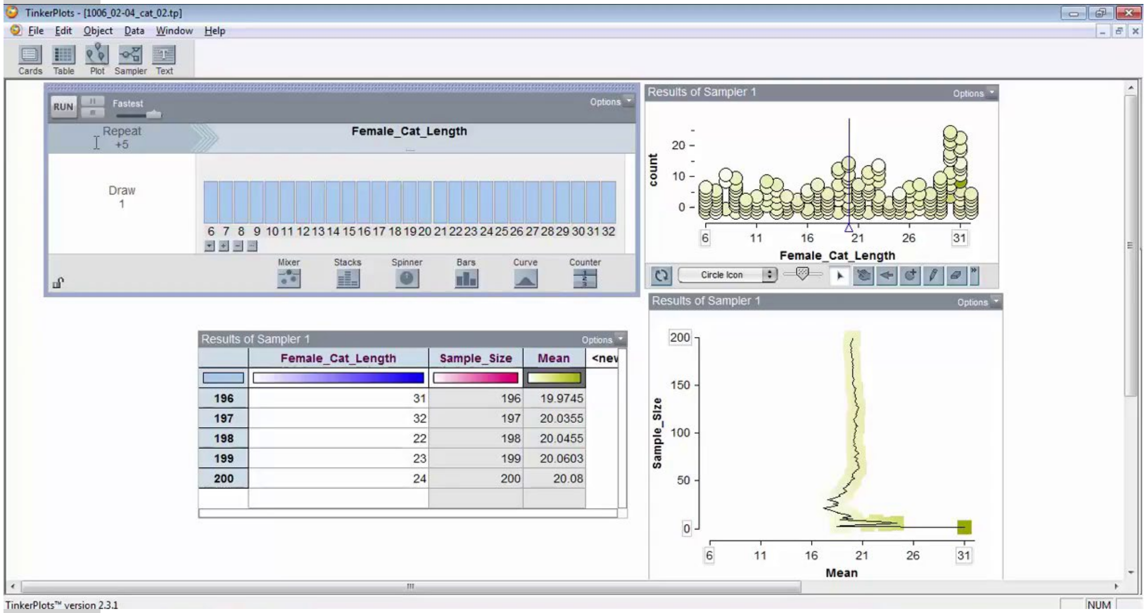
2.1192:
2.1193: S: Um, I guess the mean -- like I say, if all equally
2.1194: likely chances, so.
2.1195: Maybe.... I don't know.
2.1196: [00:45:51]
2.1197:
2.1198: I: And can you describe what you just drew
2.1199: there?
2.1200:
2.1201: S: Um, well I made the zig zag, where, like,
2.1202: each end of the zig zag, like, more likely
2.1203: to go, like, on each side. That -- so, like 32
2.1204: and 6. Um,

2.1205: I think the mean can kind of vary
2.1206: cuz if every single chance is, like,
2.1207: likely, like, you could potentially have,
2.1208: like, a lot of runs that you get, like, a
2.1209: really large cat, compared [Hm] to like a
2.1210: [00:46:16] really small, just for like the --
2.1211: trial. So I think the mean could
2.1212: definitely, like -- it'll var-- I think
2.1213: it'll -- it'll vary, and it probably won't
2.1214: be centered in one area -- area, maybe. [OK]
2.1215:
2.1216:
2.1217: I: So say that again, It'll be -- it'll vary
2.1218: and it won't be centered in one area?
2.1219:
2.1220: S: Like --
2.1221: while, like, the other ones -- they, like, were
2.1222: usually, like, centered around, like, say,
2.1223: like, say 20, or something, like, it won't
2.1224: [00:46:42] really have, like, one spot that'll
2.1225: probably, like, kind of stay in
2.1226: potentially. Like, [OK] there could be
2.1227: like a quite a few rounds that where, like,
2.1228: the average is like more like on there
2.1229: like towards thir-- 32 but it could also like
2.1230: go back to a different spot. [OK] So.
2.1231:
2.1232: I: And, uh --
2.1233: what'll happen as the sample size
2.1234: increases?
2.1235:
2.1236:
2.1237: S: Um, I think the zig zag will get smaller --
2.1238: [00:47:11] just like -- it -- the -- like the range of it won't be
2.1239: as big, but [OK] -- just because like it'll
2.1240: change like the average overall to be
2.1241: different. But, um, yeah. I'd probably just say
2.1242: like the -- like the -- each end of like the
2.1243: zig zag probably will get smaller.
2.1244:
2.1245: I: Okay.

Um, and will it -- um, so it'll zigzag less. Will it go towards any particular part, or will it, um -- of the graph or would it just -- would it kind [00:47:43] of keep wandering even with the smaller zig zags?

S: It might center -- I'm not quite sure. It might center around the 50 -- like fi-- like, I guess, like, whatever the m-- like, the middle of these two are because, if it's going back and forth a lot. But if it's not, like, it might like go off towards like one area and wander more. [OK] So I'd say maybe wander more, like it'll -- it's not like afraid to move around. [OK] /cs/

I: All right. [00:48:11] Well, let's see what happens. [OK] So what do you notice?



Results of Sampler 1

	Female_Cat_Length	Sample_Size	Mean	<nev
196	31	196	19.9745	
197	32	197	20.0355	
198	22	198	20.0455	
199	23	199	20.0603	
200	24	200	20.08	

S: It centered around, like, 21ish. [OK]

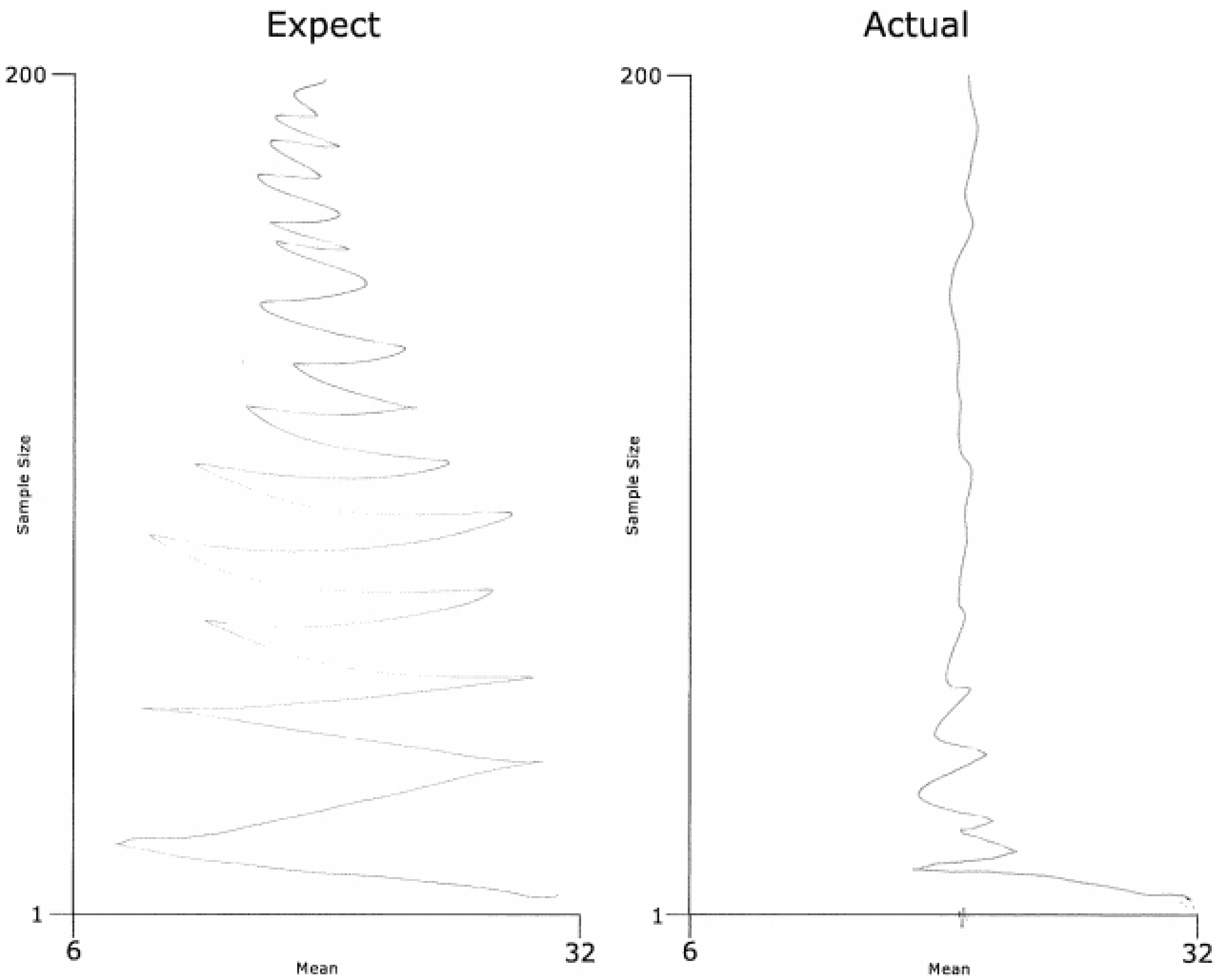
Yeah. So it did center. But -- and it didn't really deviate once, like, the -- like it never really went towards like the six area. It kind of just like stayed higher. Like, it started out 31, and then it went down, and then it kind of just like kept on like staying like near the 21.

I: [00:48:46] Okay. And so why do you think that would be?

S: Um, I guess -- it looks like in the very beginning, there was probably more of the higher numbers for the mean, maybe. Um, and so then, it kind of, like, made it be more around that number. But it kind of -- I don't think there's a lot of -- like, ju-- say like in this one, like, I don't think there was a lot of the lower numbers that, like, changed it quite as much in [00:49:18] the beginning. [Hm] So by the time, like, say, like, there were lower numbers, then, like, the -- I guess like the average between the two. So like if you have, like, in like the sample of, like, the five or whatever, it'll be, like, 31 and six, like, the middle of that is probably going to be more towards like the twenty area, say, than, like, centered around somewhere else, so.

I: Okay. Um, can you draw what you [Yeah] actually saw? [00:49:54] And so, uh, what do you notice that's similar to the previous graph?

2.1306:



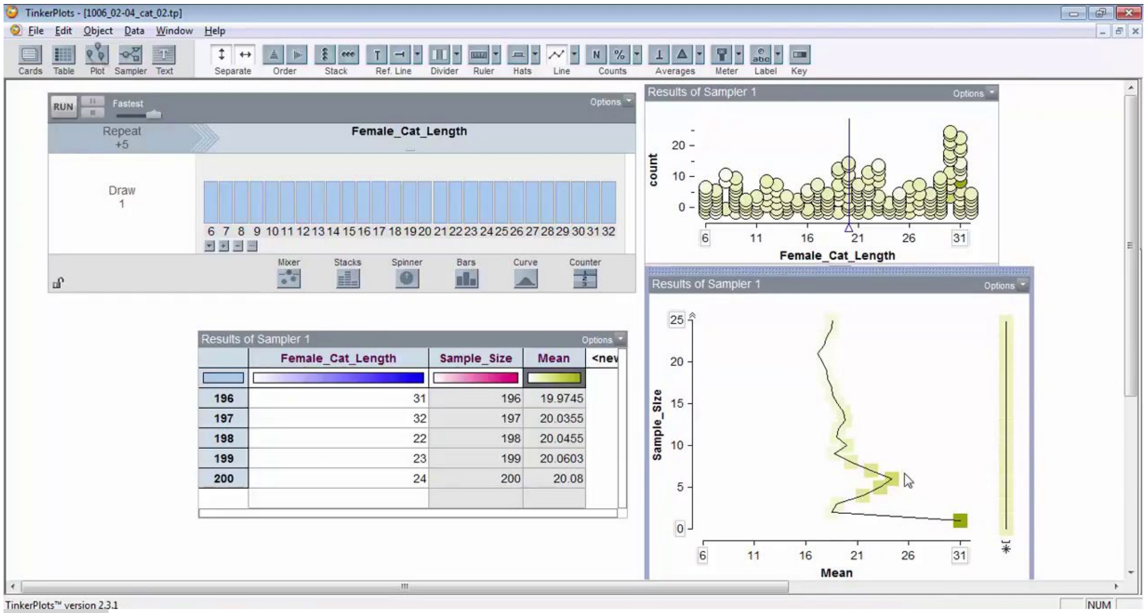
2.1307: S: Um, I guess, like, the
2.1308: zigzag at the bottom like how it was
2.1309: very like drastic, and then it kind of
2.1310: just, like, quickly tapered off, like, and
2.1311: kind of -- or not tapered off, but, like, it
2.1312: kind of quickly like stopped zigzagging
2.1313: so much, and went just kind of straight up.
2.1314: [OK] So.
2.1315:
2.1316: I: And what's different between the
2.1317: two graphs?
2.1318:

S: Um, I would say that ... it -- it seems [00:50:28] to like s-- have centered a -- much later, I guess, um, for, like, the other ones, a lot of times, like you know, take, like, a couple trials and then it pretty much stayed the same. [Hm] While this one it like went back and forth and then back and then it took a little bit of time, like quite a few trials to like get up. Cuz I think it was like around like twenty or thirty it started kind of like centering and it [00:50:50] was just like a little bit more zig zaggy and then this one's like it was towards like fifty where it really started being more even and like less zig zaggy. [mh]

I: And what would happen if we -- did I already ask this? -- what would happen if we keep going on this one?

S: Um, I'd think it would start to be a pretty straight line [mh] towards prob-- I think it would stay it, um, around the twenty one area.

I: Okay. So I'm gonna zoom in one part of [00:51:18] this graph a little bit. {Ending at 25.} Just look at the first 25 trials. So, you know, I notice looking at this that there's some kinds of -- kind of spikes in the graph. [mh] There's one here on the right, and another one here on the left. Can you describe what's happening at those spiking moments?



Results of Sampler 1

	Female_Cat_Length	Sample_Size	Mean	<nev
196	31	196	19.9745	
197	32	197	20.0355	
198	22	198	20.0455	
199	23	199	20.0603	
200	24	200	20.08	

S: Yeah. So the -- the mean for that area is kind of like -- it [00:51:48] probably had a much -- very -- like, it had to either higher numbers or, like, lower numbers for, like, the numbers that it took the mean from. [OK] Um, so, like, it was kind of going back and forth more.

I: Okay. And what happens to those spikes, um, as the sample size increases?

S: Um, I'd say it's like less of the range in between the two. So, like, this one it's, like, say, like, 25 and, like, 18. While probably up higher it's [00:52:18] probably gonna be, like, say twenty and twenty one or twenty one and, like, 23 or something like that. It'll be, like, much smaller in between like distance in between the two peaks.

Wallet analogy discussion

I: Okay.

So just one other question. [mh] Um, so you brought -- a couple -- I think you brought this up the last, um, interview too. You brought up this analogy about, um -- about the money, if you have ten dollars in your wallet [Yeah] or a [00:52:55] thousand dollars in your wallet. I was just curious, um -- is that an analogy that you came up with, or were -- sort of what's, uh -- how did you -- where does that analogy [Yeah] come from =for you?=

S: =I guess= I just kind of like think about, like, if a person is like super rich, them giving, say, like or donating, like, $10 or $100. Like, it's not gonna really, like, affect them drastically. [Hm] Like, for -- while, like for me [00:53:21] being a college student. Like, I don't really have that much of an income, [Hm] and I don't work that much, so like even spending, like, a m-- like, buying a meal, say, like, for $8.00, like, I have to budget [Hm] that in, and I have to really think, like, is this worth it, like will I still be able to pay my bills, if, like, I get this $8.00 thing [Hm] while -- and like I would love to donate but, like, looking at it overall [00:53:42] it's like if I only make this much money, and I donate, like, say monthly [mh], it's gonna be much more difficult for me to do that, and, like, it's gonna make -- be like change -- m-- like, my budget much more drastically, say, than, like, a person that is like a doctor and making a lot of money. [Right] Like, them donating like say, like, ten dollars a month, or like going to a nicer restaurant like once a [00:54:06] week. It's gonna be a lot easier for them, and they don't really have to think about it as much. So. [mh] Yeah. [OK]

2.1420: I: So that -- in that you
2.1421: sort of were thinking about it in that
2.1422: context [Yeah] already. [mh]
2.1423:
2.1424: S: Or, like, losing something, like,
2.1425: some people, like, it's like -- if they,
2.1426: like, some cash on the ground, or like
2.1427: they lose it, like, somewhere in their
2.1428: room, they're gonna be like oh I think I
2.1429: had this much, while if you don't have
2.1430: [00:54:28] that much -- if you never carry around cash
2.1431: or something, like, you're gonna notice
2.1432: that difference, [mh] like, much easier. [mh] So.
2.1433:
2.1434:
2.1435: I: Okay. That's all for today. [OK] Thank
2.1436: you. [Yeah]

Appendix D3: The Mystery Mean & Growing More Means

3.1: Mystery Machine #1: One Sample

3.2: I: ... make TinkerPlots go a little more smoothly when we're using it. /cs/
3.3:
3.4: S:
3.5: [00:00:04] Yeah. Makes sense.
3.6:
3.7: I: Ch -- ch -- ch... um, so, while we're setting up
3.8: here, um, we're gonna go back to
3.9: the casino problem, um, that we talked about
3.10: in the first session, and I'd like for you
3.11: to read through the problem just again.
3.12: It's the same problem. [OK] And again just
3.13: give me your thoughts on how you would respond.
3.14:
3.15:
3.16: S: OK, so it says you work at the state casino regulation committee -- committee. Your job is to ensure that consumers are
3.17: accurately reporting to customers the
3.18: average winnings from slot machines.
3.19: [00:00:46] Suppose one slot machines pay about zero

dollars, one dollar, or twenty dollars on each game, and the machine claims that the average payout is ninety cents. You can pay -- you can play the slot machine as many times as you want, but it costs money each time. Construct a proposed strategy for determining whether the slot machine's claim is accurate.

S: So what I would do is have, um, I would probably run [00:01:16] it you know quite a few times to at least get like a good average, so not, like, only doing it like ten times, [OK] like making sure I play it, like, at least, like -- I would -- like, more than -- definitely more than ten. Like, maybe like, twenty, maybe fifty. [OK] Um, the more the better, really. Um, and then I would just, um, see what the average payout was, so just divide by the -- however many games I played, um, from like [00:01:44] the amount that I, um, won. [OK] So.

I: And, uh, let's start with a -- start with a number. So how many times would you play the game to determine your average?

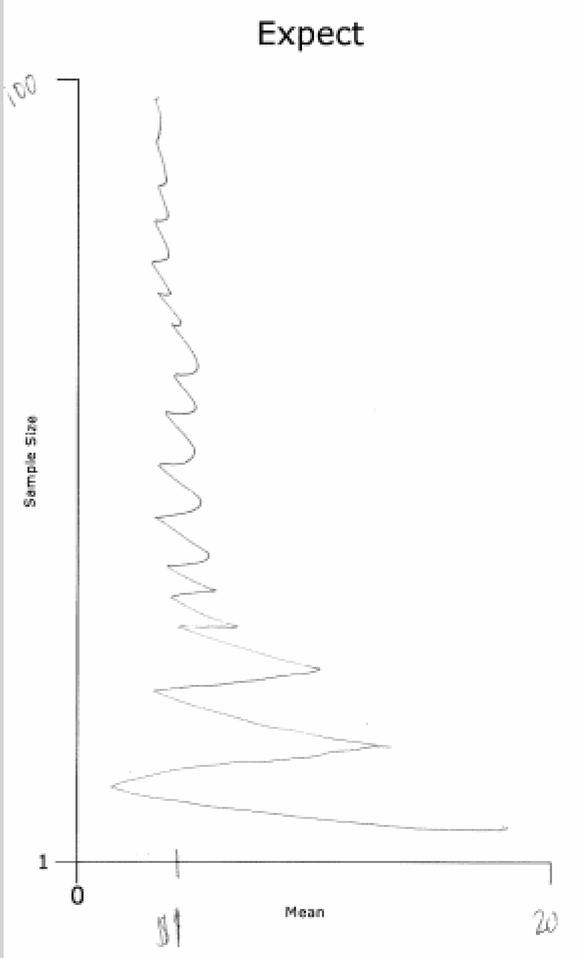
S: We'll say -- why not a hundred times. /c/

I: Okay. All right, um, so here in TinkerPlots I've set up a couple mystery slot machines. Um, so just like in the problem it gives 0 1 or 20, but we don't know how [00:02:20] likely each of those are. So that's just kind of like -- [OK] the -- like the problem. And, um, we're gonna explore what happens as we sort of take -- play more and more games to that average payout. [OK] Um, so to start out, um -- so you chose 100 -- to -- k-- as our kind of starting point for the game. So you can write 100 at the top of these graphs. And I'd like for you to sketch -- you know this is the mean against the sample size plot, so like [mh] [00:02:56] what we were doing last time, where we had the graphs going up. Um, what do you think might happen, um, to the mean going from the sample size of 1 all the way up to 100?

S: Um, yeah. Um, I would probably think -- kind of thinking about last time, like, if you say, like, win like $20 on your first time, like it's gonna be over here, but then if you win like zero time -- or $0 the next time, like it's kind of gonna go [00:03:23] zig -- it's gonna zigzag more at, like, widely I guess. Like a bigger range, like two points, um, in the very be -- like beginning, but then as you do more of the, um, slot machine then it should kind of at least the mean might go somewhere more towards, like, I guess 0 1 or 20, not really sure which one but. /c/ [OK] Yeah.

I: Okay so like for you to sketch, um, sketch what you think might happen, and this is assuming [00:03:55] that -- so let's assume that it's 90 cents for now [OK]. That the average payout is 90 cents.

S: So then, at this, like 20 time -- this could be like one dollar -- one dollar. And I guess just...



I: Okay. Um, and can you just describe what you just drew?

S: Yeah. So I did, like, a bigger zigzag at the very bottom, and then it kind of getting centered on like around ninety cents.

I: Okay and so why is -- why -- why does it look like that? Or why do you think it's [00:04:33] =like that?=

S: =Um,= because in the beginning you're gonna be -- have a bigger -- bigger variation. Like each time. Like you win twenty one

3.114: time, and then zero the next time, like
3.115: the graph is gonna move a lot more than
3.116: if say if you've done fifty ti-- like
3.117: if you've done it fifty times, um, and you do
3.118: that. It's not -- just not gonna affect the
3.119: overall like result as much.
3.120:
3.121: I: Okay. So
3.122: let's, uh -- let's see what actually happens. [OK] So
3.123: [00:05:02] you can draw a couple samples just like
3.124: before.
3.125:
3.126: S: So then I just keep going?
3.127:
3.128: I: And then let's -- uh, let's plot just
3.129: like we did yesterday the mean against
3.130: the sample size.
3.131:
3.132: S: Okay, so, uh, mean on the X --
3.133: like this one. [Yeah] OK. There. Do I have to go the other way?
3.134:
3.135: I: Yeah, you have to go up. /cs/
3.136:
3.137: S: {Oops.}
3.138: There we go.
3.139:
3.140: I: And then drag out the mean as
3.141: well. [I--] Oops.
3.142:
3.143: S: Okay.
3.144: OK. There we go.
3.145:
3.146: I: Just give you a little TinkerPlots
3.147: tip. So when it's like this /with bins/, you
3.148: can actually drag -- cli-- if you actually get
3.149: [00:05:49] one of the bars, then you can just click it. [Oh.] Um.. that didn't work quite as well as I thought. [/c/] Oh, I
3.150: drag -- did I drag -- oh I dragged other way. Okay.
3.151: [OK] So yeah.
3.152:
3.153: S: /c/ =That's helpful.=

I: =I don't know if that's= helpful. But it's a little faster =sometimes.=

S: =/c/ In the future.=

I: You can try that out. /c/

S: Okay. So then, should I do it, like, 100 times, or...?

I: Um, and so let's do -- change a couple of those options again. So let's go from -- to the borderless icon. [OK] Just make it easier to see. =And turn on the line.=

S: =And then the line.=

I: Yeah. Okay. [OK]

S: And then start?

I: Yep. /range is narrow and many are showing up as missing values/

S: Oh no. =What happened?=

I: =I think...=

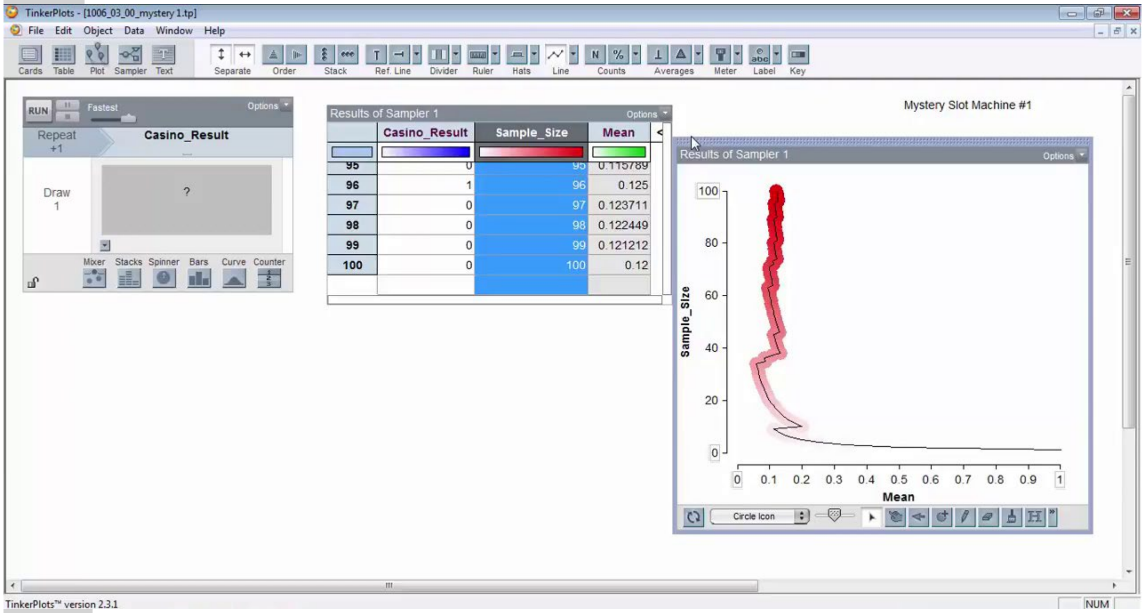
S: Should I change the --

I: I think change that one.

S: Maybe to like [00:06:30] 20? Because I guess I can't go higher than 20. = /inaudible/ =

I: I think it's going... let's just recreate the plot. [OK] I'm not sure what happened, I'm sorry.

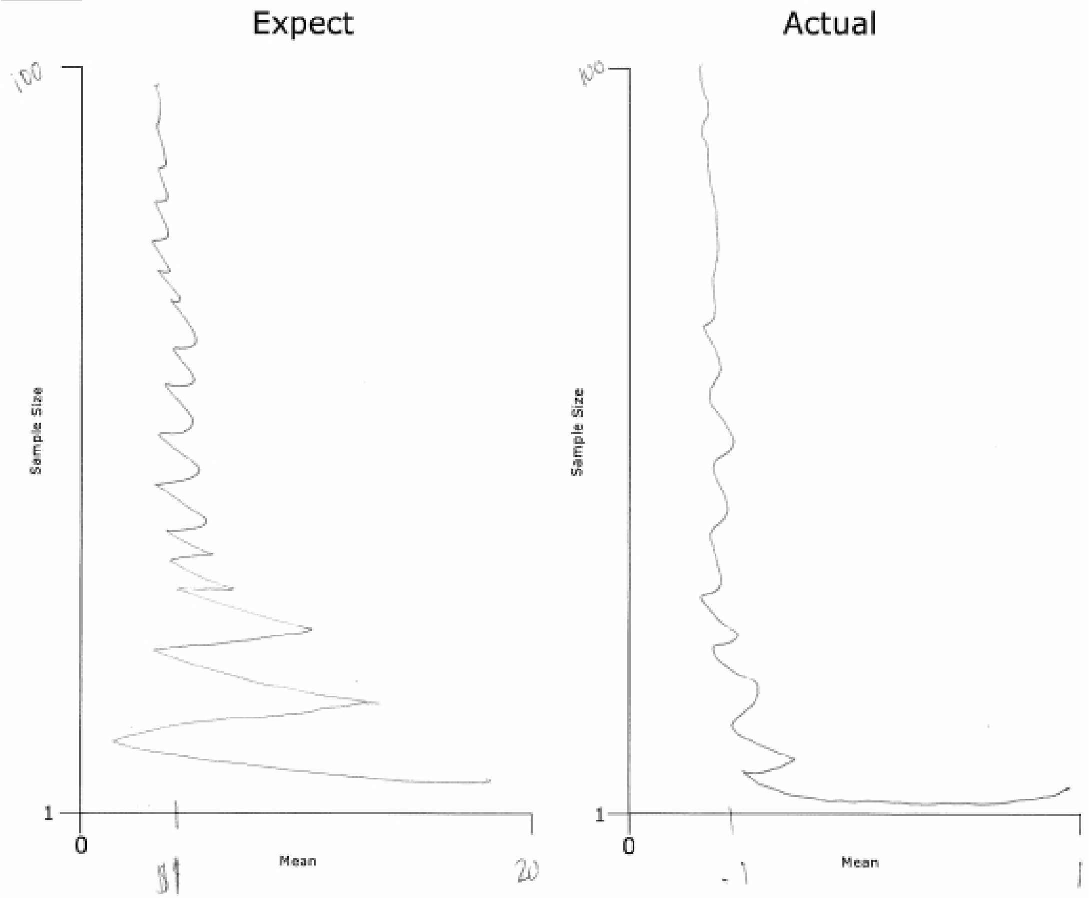
3.195: S: No, we're good. Um, delete plot. OK.
3.196: See if it works. /drags binline/ Haha! There we go. Um. [OK] Should we -- let's see, borderless. [OK]
3.197: I think it [Great.] works now! [/c/] {OK. /inaudible/ OK.} OK.
3.198:
3.199:
3.200: I: Okay. So what do you notice so far?
3.201:
3.202: S: Um, it
3.203: was more of like -- it's more a like curve than
3.204: like zigzagging all the time. [OK] Um, I
3.205: think it's probably because it's taking
3.206: like the mean each time, so like it's not
3.207: [00:07:35] gonna vary quite as much, say, than,
3.208: like, each individual like time you go.
3.209: [OK] So.
3.210:
3.211: I: Um, can you make it a little bit
3.212: bigger, just so [Yeah] we can see. And -- so you can draw
3.213: what you actually saw [OK]. You can kind of
3.214: put the axes wherever makes sense to you.
3.215:


TinkerPlots - [1006_03_00_mystery 1.tp]
File Edit Object Data Window Help
Cards Table Plot Sampler Text Separate Order Stack Ref. Line Divider Ruler Hats Line Counts Averages Meter Label Key
RUN Fastest Options
Repeat +1
Casino_Result
Draw 1
?
Mixer Stacks Spinner Bars Curve Counter
Results of Sampler 1 Options
Mystery Slot Machine #1
Results of Sampler 1 Options
Sample_Size
100
80
60
40
20
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Mean
Circle Icon
TinkerPlots™ version 2.3.1
NUM

Results of Sampler 1

	Casino_Result	Sample_Size	Mean
95	0	95	0.115789
96	1	96	0.125
97	0	97	0.123711
98	0	98	0.122449
99	0	99	0.121212
100	0	100	0.12

3.216: S: {OK, so, like, 1 is there. And then, I guess...}
3.217:



I: And, um, so how do you -- what do you think so far about whether the machine's claim is accurate.

S: I'm probably wouldn't say it's accurate, just like looking off of 100 [00:08:27] trials, cuz immediately it went down to, like, only like, um, like ten or twenty cents, like somewhere in between there, instead of ninety, like ninety cents. Like that's a -- quite a bit of a difference, I would say. Lke, [OK] but especially because it went all the way down there and stayed down there from like right in the beginning, so. [mh]

3.235: I: Um, okay. So based off of this
3.236: you're thinking that it's inaccurate. [mh]
3.237: [00:08:55] Um, how comfortable do you feel with that.
3.238: Do you feel pretty confident? Do you feel
3.239: like you have good evidence that it's
3.240: inaccurate? Or do you feel like you need
3.241: more evidence?
3.242:
3.243: S: I'd say just doing, like, a
3.244: hundred trials, I'd say I'd be, like,
3.245: comfortable saying that it's probably
3.246: not accurate. Cuz if most -- most people
3.247: if they play they're only gonna pay like --
3.248: play like three times, or something like
3.249: [00:09:18] that. Like, they're not gonna be playing
3.250: probably like a hundred times in a row.
3.251: But in general I think that it's
3.252: probably inaccurate. /c/

3.253: Mystery Machine #1: Many Samples

3.254: I: Okay. So, um, we're
3.255: going to be going back to kind of a
3.256: concept that you talked a bit about in
3.257: class, and then we did a little bit on
3.258: the first day with the post office [OK]
3.259: problem, where we're going to be looking at
3.260: the distribution of means. So this is when
3.261: [00:09:45] we collected 100 games just one time. So
3.262: this is kind of one trial that we just
3.263: did. But if we did many trials so if we
3.264: plot this, take the mean, collect
3.265: statistic, and gather many means,
3.266: um, at --
3.267: so our sample size is a hundred here. So
3.268: could you -- you [Yeah] can write a hundred up there. Um.
3.269: What do you think the distribution of
3.270: those means would look like? So if we did
3.271: [00:10:16] this over and over again. So this time we
3.272: got 0.12, um, but if we did it over and over
3.273: again what do you think the -- the distribution

of those means would look like?

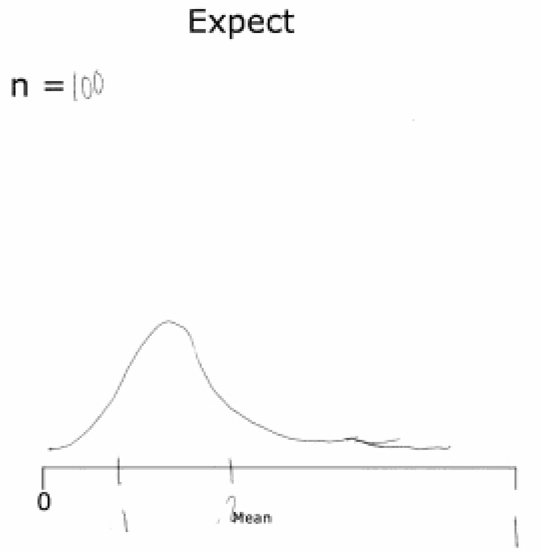
S: Um, I would say it probably would stay somewhere in between, like, point one and point two [OK], um, over time. Um.

I: OK. So can you sketch a graph do you think that would look like?

S: Cuz I think I'm gonna -- just, like, make this one. So then -- is it just kind of like -- OK. Is it kind of like the one before, or...

I: So [00:11:01] it's not gonna be in this kind of thing, it's gonna be each -- you know how in TinkerPlots you have each dot as a mean [Yeah], and then you have some kind of shape of [OK] those means overall. Yup.

S: So I'd probably say, like, kind of just, like, going down, maybe. Does that seem -- [OK] Yeah.



I: So we're actually gonna do that. We can [OK] actually do that fairly easily in TinkerPlots. So the first thing we can do is if you can drag down -- um, if you can drag the plot of these 100 means.

S: [00:11:33] Okay. So then just, like, make a new plot? =Or--=

I: =Um--= Well actually, let's change the sampler first. [OK] So, um, we want to change it -- just like how we set it up in -- in 3264, usually. Um, so our sample size is 100, so we change the repeat to 100. And right now it's set with that kind of plus sign on, so it keeps [mh] adding on. So you want to turn that back to normal by going to --

S: = /inaudible/ =

I: -- Sampler Options, and then Replace Result Cases. [OK] [00:12:01] And it then if you run it once, it'll be -- we'll get a different sample, um, and a different graph there. [mh] Um, so we can drag down -- so let's look at the plot of the results, the casino result.

S: OK, so then, new plot -- here [Mh, yup], OK. And then -- is the X ax-- the X-axis fine?

I: Sure.

S: Okay.

I: We can stack just to see what that looks like. We can look at the mean. Um, so that time we got point three three. Um, so let's collect the statistic.

S: Of the [00:12:41] mean?

I: Yeah. [OK] And let's -- let's do it total of 500 times.

S: OK, I'll make everything small. [/inaudible/] /c/ Does it matter if I close this one, or not really?

I: Um, that one doesn't matter actually, yeah. [OK] Because anything that's of the original sample will slow it down. [OK]

S: And then how many times did we say?

I: 499. [OK]

S: OK.

I: OK, so let's see what actually happened.

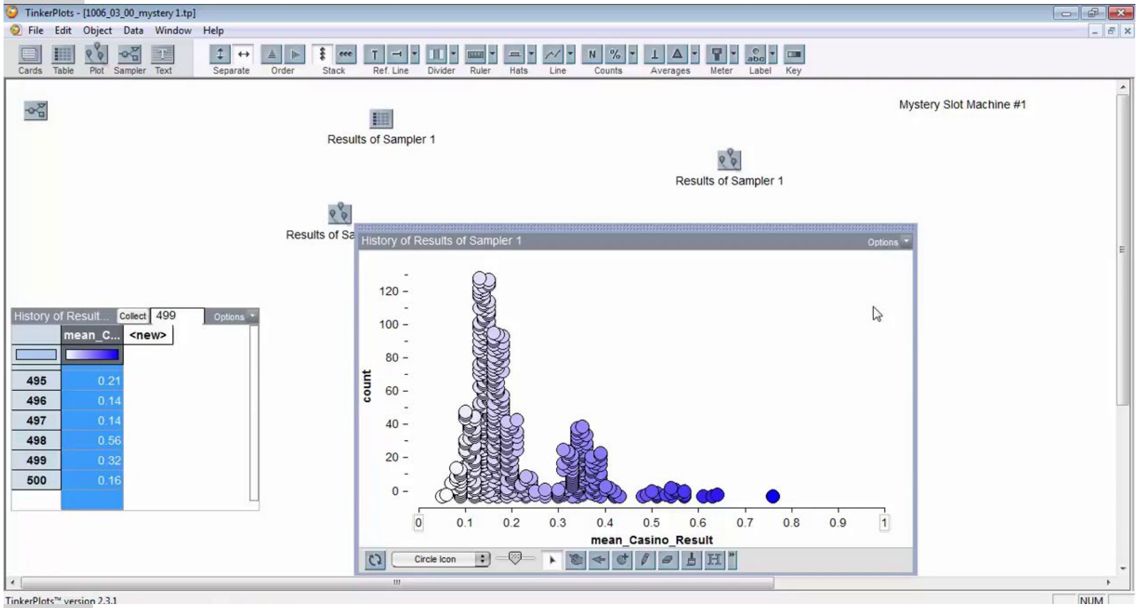
S: =Isn't it...=

I: =So that's the original= sample. [Oh yeah] That's -- o-- just one sample. [OK]

S: So then I'll a new plot [mh], or... OK.
[00:13:39]

I: So maybe you can set the end axis to one, just so it matches [OK] up with... [OK] And so just describe what you see for me at first.

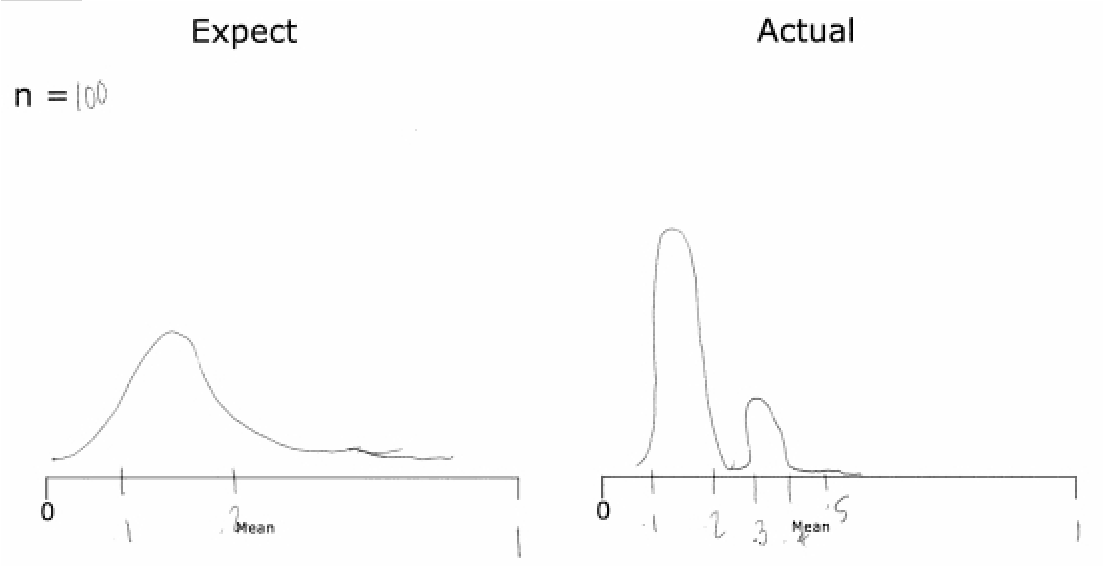
S: Yeah, so it looks like there's kind of, like, two little, I guess, like, are they called, like, modes, or like bumps in the graph. So there's one. It looks like in between point one and point two, and then there's another one in between like point three and point [00:14:08] four. But the one, um, in between point one and point two is much larger than the one in between point three and point four. [OK] And then it definitely, like -- it doesn't really have anything past, I'd say, like, point, like, sixish.


TinkerPlots - [1006_03_00_mystery 1.tp]
File Edit Object Data Window Help
Cards Table Plot Sampler Text Separate Order Stack Ref. Line Divider Ruler Hats Line Counts Averages Meter Label Key
Mystery Slot Machine #1
Results of Sampler 1
Results of Sampler 1
Results of Sa
History of Results of Sampler 1
Options
count
0 20 40 60 80 100 120
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
mean_Casino_Result
Circle Icon
History of Result... Collect 499 Options
TinkerPlots™ version 2.3.1
NUM

	mean_C...	<new>
495	0.21	
496	0.14	
497	0.14	
498	0.56	
499	0.32	
500	0.16	

I: Okay. Great. So, um. Alright, so you can draw that. [OK]

3.394:



3.395: S: Let's see. /If I?/...
3.396:
3.397:
3.398: I: Um, and, uh, does this -- is this what you expected,
3.399: or is it different than what you
3.400: expected?
3.401:
3.402: S: It -- I didn't expect there to be
3.403: [00:15:02] two. I expected it to be more just like
3.404: one single, like, bell curve [mh] instead of
3.405: two small -- or two of them. So. [mh] But I did -- I
3.406: guessed that it'd be around, like, point
3.407: one, point two at least. [OK] At least in this case
3.408: for one of them, so. [OK]
3.409:
3.410: I: Um,
3.411: and so how -- how -- what -- percentage of the
3.412: time did you get results that were 0.9
3.413: or higher?
3.414:
3.415: S: Um, 0% /c/
3.416:
3.417: I: OK. /cs/
3.418:
3.419: S: Nothing's there. /cs/

I: So based on this, do you think that -- you know, so you had just one shot [00:15:41] before. So when you actually play the game, you'll only be able to do it once, you won't be able to collect 500 trials [mh] or anything. Um, do you think that based on this, and the percentage of the time you saw point 9, do you think that you'd be able to, um, tell whether the machine's accurate or not by playing it just a hundred times?

S: Um, I would think so. I mean, each one of those dots represents the [00:16:09] mean of 100 trials [mh], so there's a lot of trials on here and the fact that none of them -- there was only one above, like, point, like, say six five -- it's really like probably -- even been playing it 20 times, like, you probably wouldn't even maybe get to, like, point 9 [OK], so. [OK] Yeah.

I: Um, so how could you be even more sure that the, um -- even more kind of confident that you have the right -- that the machine is wrong, [00:16:43] to gather even more evidence.

S: Probably just running more trials, I would say. [OK] There's not -- it never hurts to have more trials rather than less, so.

I: Okay. And so, uh, how many -- um, how many more, or...

S: Um, I'd probably keep it at like 500 for like collecting the statistic, but maybe for like the draw on the original sampler, uh, maybe increase that to, like, 200 or 300 or [00:17:13] something like that. [OK] Um, so.

I: So, let's pick a number.

S: Okay, we can do -- we'll do 200. [OK] =/inaudible/=

I: =And so= before you do that, [OK /c/] I just want to ask. So why -- why would you be more sure with 200 than you would be at 100.

S: It's twice as many people. Um, each, like, each time you get the mean and then when you collect the t-- statistic it's gonna be twice as many, um, for just like the mean average, so.

I: Okay. Okay, great. So let's see.

S: So I think it's this one? Yeah. OK. So then [00:17:48] 200. And then should I --

I: So you can just shrink this down. It's already --

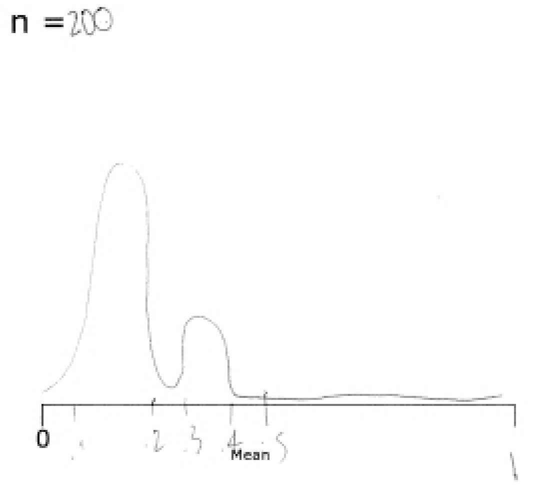
S: Okay.

I: And then [And then --] you can just =delete these.=

S: =Delete the case.= [Yup.] OK, yeah. Um, where is it?

I: It's down there a bit -- oh maybe. I think you have to click on the corner. [Oh, yeah.] That one. Yup.

3.502: S: There we go.
3.503: Delete All History Cases, and then...
3.504:
3.505: I:
3.506: So you can change that to 500. And so let's let
3.507: that -- we're gonna let that start running,
3.508: [OK] but I'm gonna minimize it /c/ [Yeah]. Um, and
3.509: [00:18:23] can, um, you draw what you think will happen.
3.510:

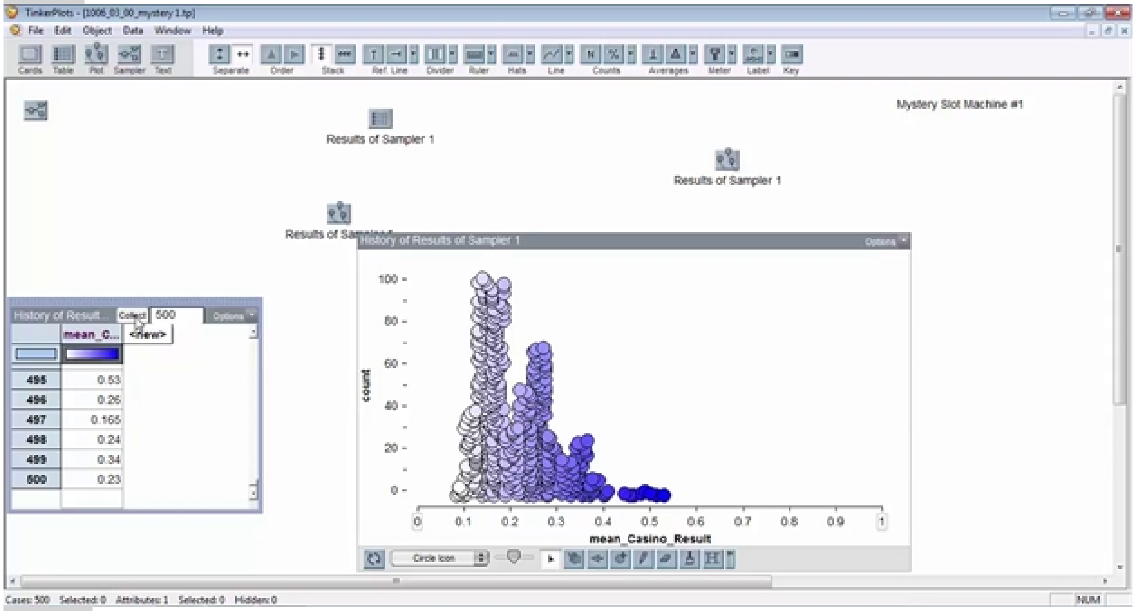


3.511:
3.512: S: Yeah, okay. So this was 200 now. [mh] Um, I feel
3.513: like it'll probably stay pretty close to
3.514: what we had before, um,
3.515: so we'll say, like, this is point five, and it'll
3.516: kind of go up and then up again and
3.517: then probably go off one more time, so this is like point two, um, point three, and point four.
3.518:
3.519:
3.520: I: Okay. And why would you expect that to happen?
3.521:
3.522:
3.523: S: Um, I -- I don't really expect that much of a
3.524: difference. Just because, um, we had so many
3.525: [00:19:04] trials in the previous one as well. [mh] So I
3.526: feel like if anything, um, it'll just have --
3.527: be, like, larger like the -- each point will
3.528: be larger, probably. [OK]
3.529:

I: And so could you describe what you mean by each point being larger?

S: Like each, um, like, bump in the graph or like when it goes up it'll just kind of go up even more because there's more like on -- like there's more trials, so. [OK]

I: OK, great. Well, let's see what actually happened.



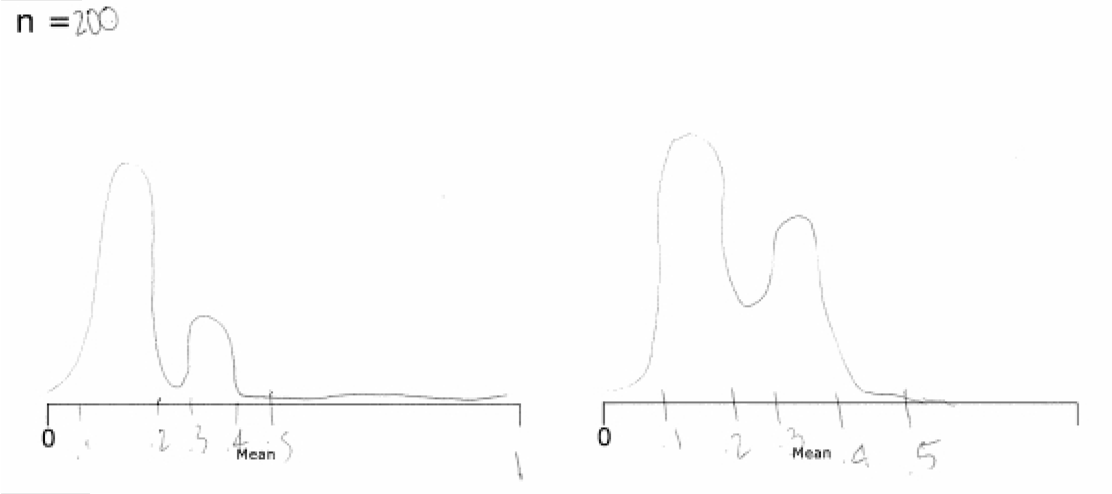
TinkerPlots - [1006_03_00_mystery 1.tp]
File Edit Object Data Window Help
Cards Table Plot Sampler Text
Separate Order Stack Ref Line Divider Ruler Hats Line Counts % Averages Meter Label Key
Mystery Slot Machine #1
Results of Sampler 1
Results of Sampler 1
Results of Sampler 1
History of Results of Sampler 1
Options
count
100
80
60
40
20
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
mean_Casino_Result
Circle Icon
History of Result...
Collect 500
Options
<new>
Cases: 500 Selected: 0 Attributes: 1 Selected: 0 Hidden: 0
NUM

	mean_C...
495	0.53
496	0.26
497	0.165
498	0.24
499	0.34
500	0.23

S: [00:19:39] So there's a lot more in between. So it's less, like, I guess drastic. It's more like combined rather than separated from the first one. [OK] But it's still -- like, there's nothing even above point five for this one, so. [OK] Or like 0.55, I would say, maybe. [OK] So.

I: Okay. So you can draw that. [OK] And just comparing it to what we saw with 100 is getting 90 cents, uh, more likely

[00:20:30] with 200 trials, about the same likelihood, or less likely =than 100 trials.=



S: =I think= it's even less likely. [OK] Because before we at least had like one point that was like around like point six or point seven. But this time we don't even have anything like above, like, basically 50 cents. [OK] So.

I: And why do you think that would be?

S: Um, probably because the -- you're taking the mean from even more trials, so it's [00:20:55] going to be -- if, say, like, getting one dollar two dollar was really, um -- or like one dollar or twenty dollars was unlikely, um, just adding a hundred more into that. Like that's more zeros just adding into that [Hm]. Like, into the mean. Um, and so it's gonna definitely decrease the -- um, am-- like mean for each like trial that we did of 200.

I: OK. And you said something about it being more combined. Can you tell me a

[00:21:24] little bit more about that?

S: Yeah. Like for the one with only a hundred trials, the two, like -- I guess, like, bumps or like peaks or whatever they're called. Um, they were a lot more separated. Like it really wasn't in betwee-- anything in between like point two and point three. [mh] But for this one, um, there's quite a bit more in betwee-- like between the two, um, well, like bumps instead of it being like closer to zero, the [00:21:50] mid-- like in-between it it's like higher up. [OK] So. And they're definitely more even in height.

I: And so why do you think that would be?

S: Um, I'd say just having more trials and, like, it's -- I'd say just like having more trials, like, it's kind of like just evening out I guess a little bit more. Um, [OK] and probably getting more centered -- or like it's just kind of -- yeah it's just [00:22:20] evening out. /c/

I: Okay. Can you tell me more about what you mean by evening out?

S: Like it's gonna -- like each dot isn't gonna be moving quite as much as like if you only have like say 20 trials. Like it doesn't make as much of an impact. [OK] So it's gonna be more likely to be like around what the average amount of pay -- payout is like in general. [OK] So. [OK. OK. Um...]

I: [00:23:00] Okay, great. So that's one possible way the machine could have worked. [OK] So we're gonna -- um, and I'm just moving these out of your way --

S: = Yeah, of course.=

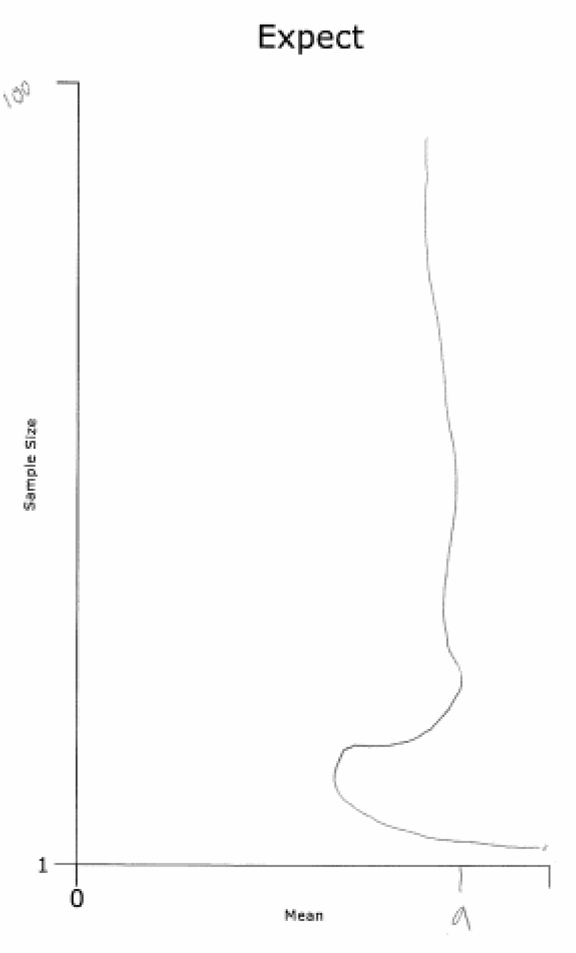
Mystery Machine #2: One Sample

I: -- but if you ever wanna look at them again, um, just me know. Um, so, um, now we've got another mystery machine. So this is also a way that the machine could be working. Um, and it's still kind of a mystery, so it's set up pretty much the same way. Um, and I'd like for you to draw again [00:23:35] -- so again assuming that they're correct, [mh] that it is 90 cents, just to draw, um, what you'd expect, um, at equals 100. Um, and if you expect the same thing -- if your expectation hasn't changed from last time, you can just draw the same graph as last time, or you can draw something different if you want.

S: So, like, the expectation of, like, it being like 90, like their claim being accurate?

I: Yeah yeah yeah. [OK] What -- what that [00:24:02] graph would look like.

S: I would probably say it's probably close to like what the actual was for that, so maybe, like, kind of ... nyuuuuh, and then it going to, like, 90 cents.

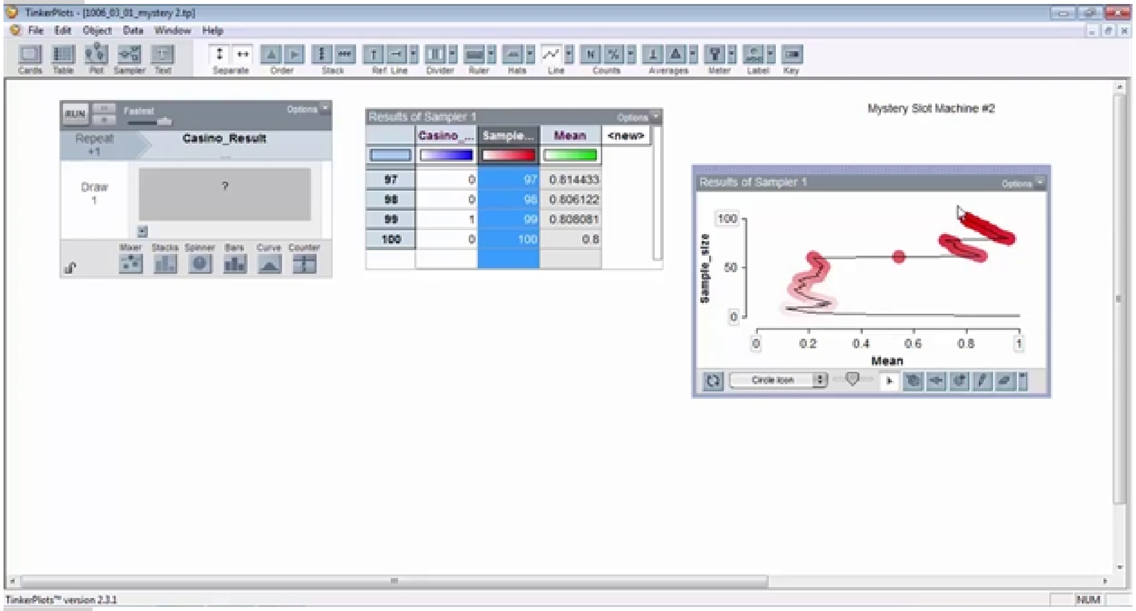


I: Okay, great. And so, um, so this is a bit different than your first one. [mh] What sort of changed?

S: Um, I didn't really, like, think about like the means, like the means are gonna be less likely to like be extremely different than, um, say like a one or two trial -- like [00:24:35] doing, like, y-- an individual trial for like -- like on the x-axis. Like if you win, like twenty dollars or like one dollar. Like if you're taking the mean of like your trials like it's not gonna be like super drastic of a difference like

3.675: each time. [OK] So.
3.676:
3.677: I: And why is that?
3.678:
3.679: S: Um, because
3.680: you're taking an average of like an
3.681: amount, instead of like say like twenty
3.682: and one, like it's gonna -- that's gonna
3.683: [00:25:01] change a lot more and like an av-- the
3.684: average of that like isn't going to
3.685: change as much each time like you have
3.686: like your own trial.
3.687:
3.688: I: Okay. Great. Okay, so
3.689: let's see what actually happens again. So
3.690: you can draw a couple and then we can
3.691: plot it again.
3.692:
3.693: S: Okay. And then...
3.694:
3.695: I: And let's draw that
3.696: mean by sample size plot.
3.697:
3.698:
3.699: S: And then, the borderless, and the line. OK. And then... /several 20s appear, mean is at 0.8/
3.700: So that one definitely had a jump. /c/
3.701:
3.702: I: Okay.
3.703: [00:25:57] And so what's that jump?
3.704:
3.705: S: Um, I'm guessing
3.706: that in, like, the result probably, um, there
3.707: was like a 20 or maybe like two in there.
3.708: Cuz it -- definitely cuz a -- a $1
3.709: wouldn't just give it, like, a point -- like
3.710: around point three of a difference. It would
3.711: probably be like the $20, um, putting it in
3.712: there. [mh] So and then it -- uh, it like changed the mean
3.713: for the later ones, too. So.
3.714:

3.715: I: Okay. Um. Okay
3.716: So you can draw what you actually saw. [OK]
3.717:



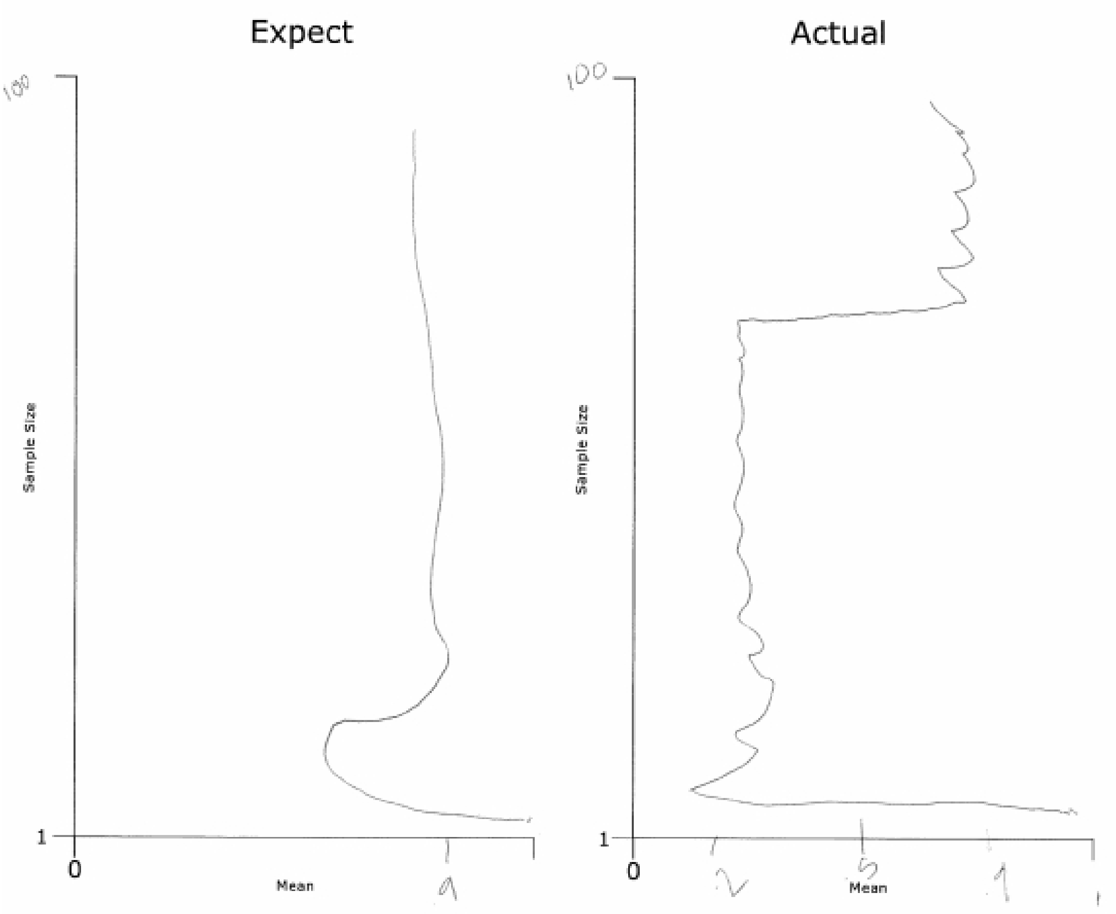
TinkerPlots - [1006_03_01_mystery 2.tp]

Mystery Slot Machine #2

Results of Sampler 1

	Casino_...	Sample...	Mean	<new>
97	0	97	0.814433	
98	0	98	0.806122	
99	1	99	0.808081	
100	0	100	0.8	

3.718: S: { /inaudible/ }

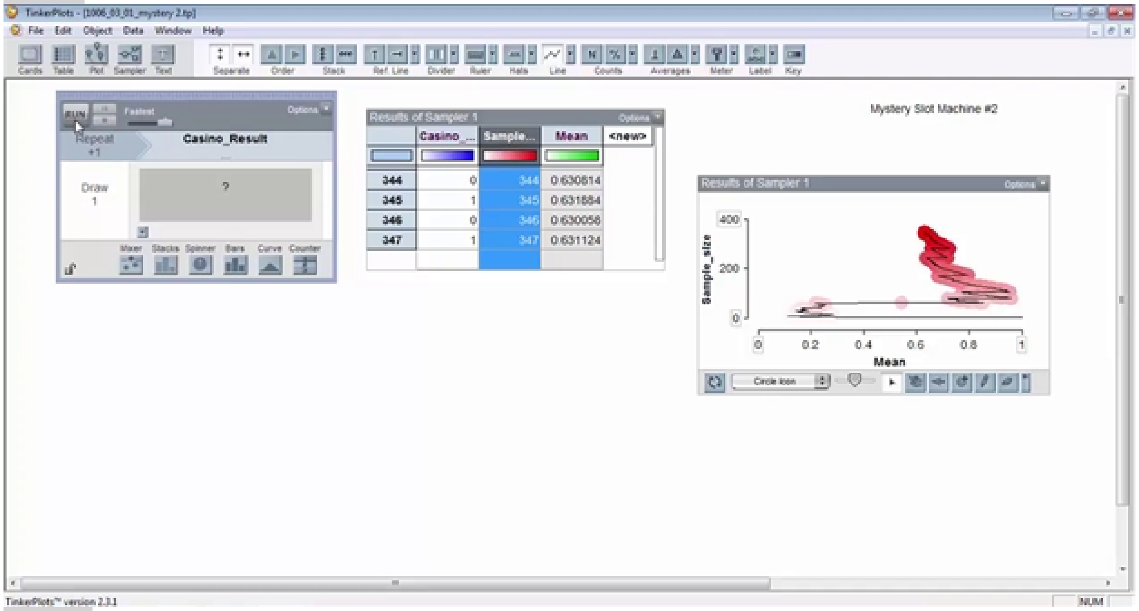


I: And [00:26:56] about -- okay. And about where did it end up at the end there?

S: Um, I'd say like pretty close to -- point 9 or so, I guess. Yeah, it's at -- say close to like -- eh. Close to like -- in between like 0.9 and 0.8 [OK] I'd say.

I: Okay, great. And how do you -- do you feel like you can -- can you tell at this point whether the machine's claim is accurate?

3.735: S: I'd say no. You definitely would have to
3.736: take a lot more samples. Because if one
3.737: [00:27:34] mean can change the plot of the graph, um, so
3.738: drastically, I would say definitely like
3.739: you would need many more samples /c/ or, um,
3.740: trials at least to give like an accurate, um,
3.741: claim on what it is, so.
3.742:
3.743: I: Okay.
3.744: Um, so let's, um -- let's just keep going for
3.745: a while. [OK] And you can stop whenever you
3.746: sort of feel satisfied that you've gotten
3.747: there. [OK] Um, that you can make a claim. So
3.748: let's just keep going
3.749: [00:28:45] for a bit longer. [OK]
3.750: /Stops at 347/ So what --
3.751:


TinkerPlots - [1006_03_01_mystery 2.tp]
File Edit Object Data Window Help
Cards Table Plot Sampler Text Separate Order Stack Ref. Line Divider Ruler Hats Line Counts Averages Meter Label Key
RUN Fastest Options
Repeat +1
Casino_Result
Draw 1
?
Mixer Stacks Spinner Bars Curve Counter
Mystery Slot Machine #2
Results of Sampler 1 Options
Sample_size
400
200
0
0 0.2 0.4 0.6 0.8 1
Mean
Circle Icon
TinkerPlots™ version 2.3.1
NUM

Results of Sampler 1

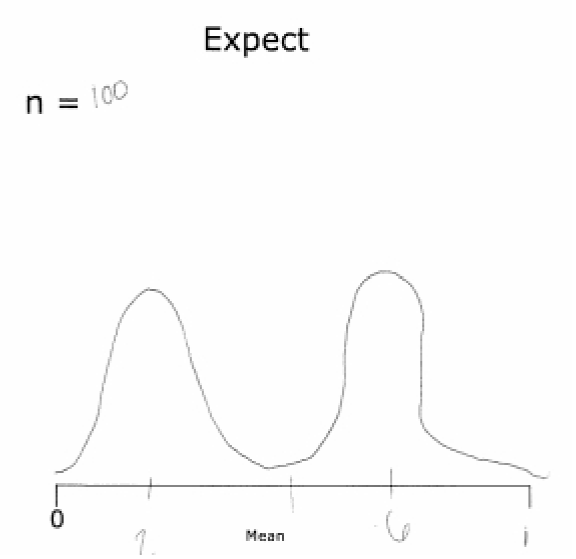
	Casino_...	Sample...	Mean	<new>
344	0	344	0.630814	
345	1	345	0.631884	
346	0	346	0.630058	
347	1	347	0.631124	

3.752: S: Yeah. I was gonna say =I feel like --=
3.753:
3.754: I: = Um, so how're you feeling?= /c/
3.755:
3.756: S: I'd probably say --
3.757: it's probably gonna be-- stay in between
3.758: maybe at least it has for quite a -- like, two --
3.759: almost like 200 trials. Probably

3.760: somewhere in between, like around like
3.761: 0.7 somewhere in that area. [OK] Between, like,
3.762: point seven and point eight. [OK] I guess
3.763: point -- I guess that one's closer to point six.
3.764: Um, so I guess point six-- in between,
3.765: like point six and point seven. [OK] What they'll be in.

3.766: Mystery Machine #2: Many Samples

3.767: I:
3.768: [00:29:17] Okay. Um, so once again, we're gonna move
3.769: to looking at the means of trials again.
3.770: And again sort of our initial kind of
3.771: look was at one hundred trials. [OK] Um, and what
3.772: do you think the means -- the -- sort of that
3.773: distribution means of a hundred
3.774: trials would look like?



3.775:
3.776: S: I would probably...
3.777: say I guess, ... { /inaudible/ }
3.778: I'd say it's probably gonna be close to point 2,
3.779: but probably having maybe two --
3.780: [00:30:06] like of the bumps again, it's gonna be, like, um --
3.781: just to like, because there -- if there's
3.782: any 20s, then it'll definitely increase
3.783: the mean, so it was, like, say like point two,
3.784: and it's like point six, we'll say. [OK] So.

I: And just what's the /inaudible/?

S: Oh, yeah.

I: /c/ OK. Um, and... oh, so one other thing. One question I have. So we expand this view a bit, um, so, you know, we already sort of identified those kind [00:30:50] of big leaps [mh], um, um, and you said that those were 20s or something big kind of coming in. Um, what -- what happens every time that we get a twenty sort of as we go -- as the sample size increases?

S: It definitely goes, like, horizontal like some percentage like in the -- like with less sample or like with a smaller sample size, it's a lot more drastic, but it kind of like gets smaller -- like i-- it -- having like say a 20 or [00:31:22] a couple of them for the mean, like, it definitely -- it just kind of like jumps straight over, and [mh] -- but as we go -- get bigger, and have more trials then it's definitely a smaller -- like a smaller jump over. [OK] So.

I: And why is that?

S: Um, because you have a small-- you have a larger sample size so it doesn't affect it as -- quite as much.

I: Okay. Um, so -- okay. So now we're gonna do the same thing. So we can -- [00:31:48] um, again we're gonna change our sampler options.

S: All right. So then, does it matter what I should do the repeat as, or should I do like, 100, or --

I: Let's do a hundred, [OK] um, just to be consistent.

S: And then taking Replace Cases off. [OK]

I: And so let's just run that once to clear out what we saw before. Okay, so that's another sample of a hundred. [Hm] Um, and let's, uh -- let's plot those and collect the statistic.

S: OK. For the casino one [00:32:20] [Yeah], is it the right one?   Okay. OK. And then was it the mean that we took the --

I: Yup, we /record?/ the means.

S: {Oops.} And then make everything small. And then 499?

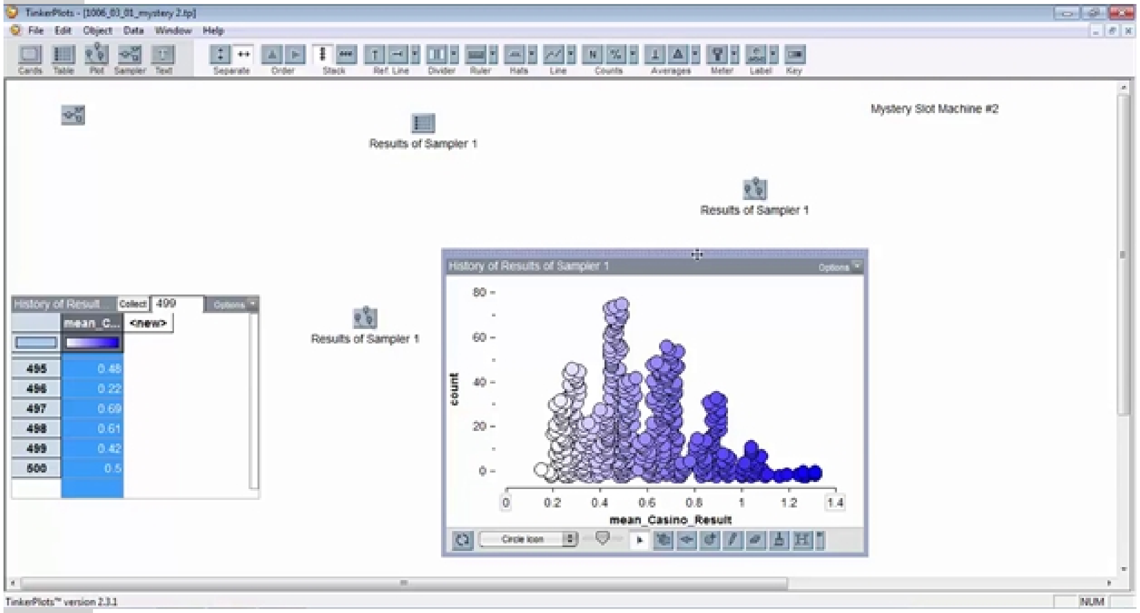
I: Yep. Um, so that's the original sample. [Oh yeah.]

S: {I keep doing that. OK.}

I: I actually learned that trick from a participant -- another participant /c/ in the study. I was like, "Ah, didn't know you do that." /c/

S: I'll take it! /cs/

3.864:

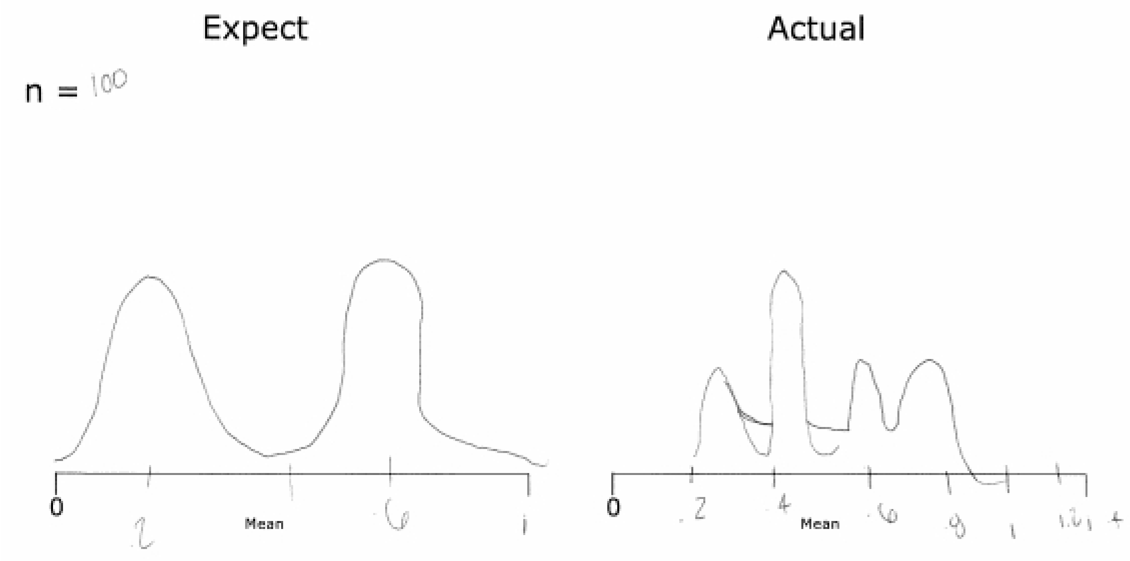


TinkerPlots - [1006_03_01_mystery 2.tp]
File Edit Object Data Window Help
Cards Table Plot Sampler Text Separate Order Stack Ref. Line Divider Ruler Hats Line Counts Averages Meter Label Key
Results of Sampler 1
Mystery Slot Machine #2
Results of Sampler 1
Results of Sampler 1
History of Results of Sampler 1
Options
count
80
60
40
20
0
0 0.2 0.4 0.6 0.8 1 1.2 1.4
mean_Casino_Result
Circle Icon
TinkerPlots™ version 2.3.1
NUM

History of Result ... Collect 499 Options

	mean_C...	<new>
495	0.48	
496	0.22	
497	0.69	
498	0.61	
499	0.42	
500	0.5	

3.865: I: Okay. So what do you notice?
3.866:
3.867:
3.868: S: Um, there's a lot more, like, areas where
3.869: it goes up and then down and up and down [OK]
3.870: [00:33:38]. Um, then in the previous one before, there
3.871: was really only like two like distinct
3.872: like parts that went up, and now there's
3.873: like I'd say like four-ish.
3.874:   [OK] So.
3.875:
3.876: I: All right. So you can draw that.



S: {OK...}
I guess they're a little bit more like...
/list?/ {Going down...}
It's like not the best drawing.

I: =That's OK.=

S: =But that's OK.= /cs/ It's the thought that
counts, right? /cs/

I: Okay. And so, um, you said
something about the shape being a
[00:34:39] little different before. [mh] Do you have any
thoughts on why that might be?

S: Um,
probably because there's a higher
percentage of being able to, like, win $20. So
there's a lot -- it's a lot more likely to
vary, um, because that definitely, like, if you
get like a zero, like, you're not gonna
make any money, but if you have a twenty
like that's quite a bit of a difference
like between zero and so it's definitely
[00:35:05] gonna affect the results a lot more. Um, so

it makes sense that, like, if you get a tw-- couple twenties in a row, like, it's gonna affect the -- the mean, um, quite a bit. So, that's probably why of just having more probab-- like a higher probability of getting a $20 [OK] in the slot machine.

I: Um, and, um, what's the -- what kind of percentage of the time would you get ninety cents or more, [/yawns/] just on a hundred trials?

S: Just, on, like a hundred trials. I [00:35:42] would say it's -- it can happen. It's a lot more likely than um, like before, but you're a lot more likely to get I would say like maybe like point five, like fifty cents. Um, instead of ninety. Like it's a lot -- it's a lot -- it's like plausible -- like it -- it could happen. Like, you never know depending on like what -- like how much you win and stuff like that. Like it -- you can't like [00:36:10] really rule out um, not like -- not being able to get ninety cents. [OK] Um as your average payout. But, you're more -- it's gonna probably lean more towards like 50 cents I would say, just looking at the graph.

I: Okay. And if you -- um, so based on -- based on this graph. So again when you actually do the study, you'll only have one shot, and you'll only be able to -- you'll only see sort of one of [00:36:39] these means. Um, do you think that somebody who played the game a hundred times would be able to accurately tell, um, what the average payout was and whether the claim was accurate?

S: Um, I don't think so. I'm just thinking about like the previous graph, um, with like 100. Like you know it kind of goes back and forth quite a bit like even with like only a hundred trials. Um, so if like you s-- get like, [00:37:04] say if you do it like twice, like if you get a zero at one time and twenty dollars the next time your average payout would be ten dollars [mh] and like that's not really like accurate so you're definitely like I'd say you need more than like a hundred trials to be able to like say that the average would be ninety, but.

I: Okay. Great. Thanks for sharing your thoughts on that, and, uh, [00:37:30] let's try-- see what happens at 200.

S: OK. So then, open this one up and put 200. [mh] {OK}

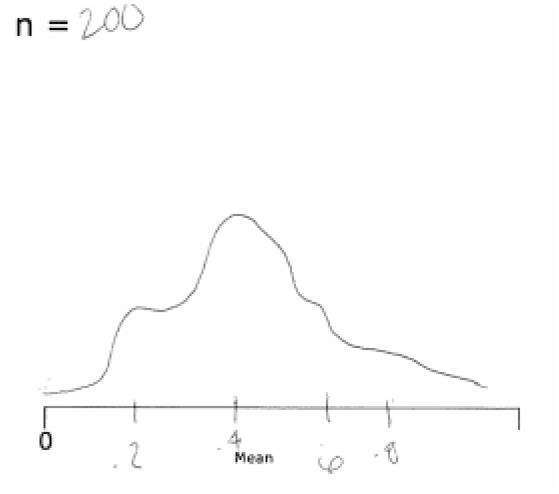
I: And then we delete those results, and then change to 500. And we can collect. And again I'll -- oops -- just -- Oh no --

S: I can...

I: Look away? [Yeah] /I minimizes TP window/ Okay, there we go.

S: There we go. /c/ =So this was 200...=

I: =So let's draw= what you think would happen, um, for 200. What you think that graph [Um...] will look like.



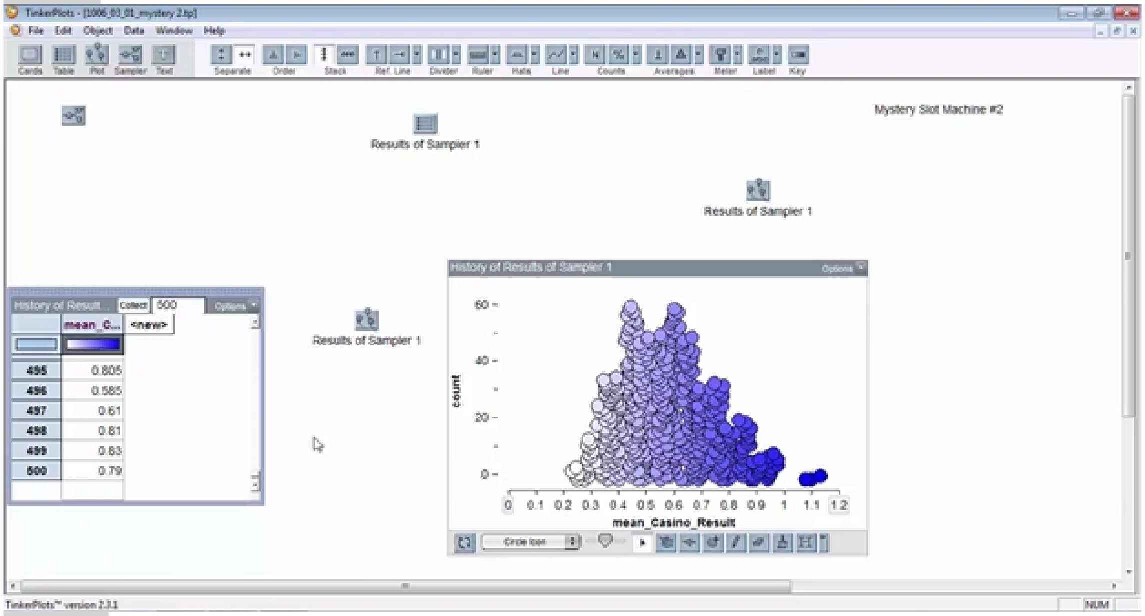
S: I feel like it'll be even like more like meshed together -- like, closer, kind of like the gaps -- like distinct like peaks [00:38:18] will become closer to each other. Um, so like... maybe?

I: Okay. And what are -- sort of what's this, can you label a couple of the values there?

S: Yeah, so I'd probably said this is gonna be close to like point fourish. Um, this could be like point two, maybe this can be like point six and point eight.

I: Okay. All right. So we can just click it to see what actually happens.

3.1001:



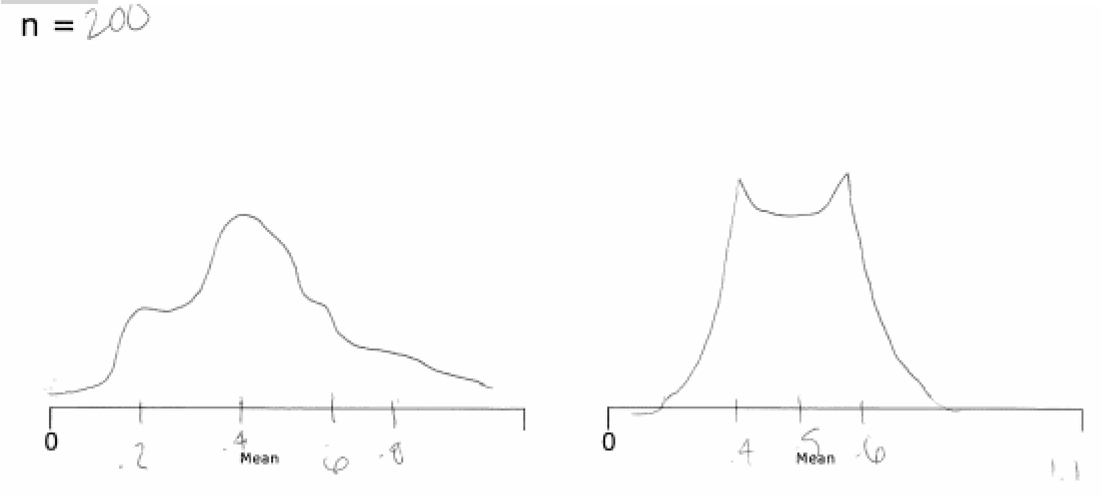
TinkerPlots - [1006_03_01_mystery 2.tp]
File Edit Object Data Window Help
Cards Table Plot Sampler Text Separate Order Stack Ref. Line Divider Ruler Hats Line Counts Averages Meter Label Key
Mystery Slot Machine #2
Results of Sampler 1
Results of Sampler 1
Results of Sampler 1
History of Results of Sampler 1
Options
count
60
40
20
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1 1.1 1.2
mean_Casino_Result
Circle Icon
TinkerPlots™ version 2.3.1
NUM

History of Result... Collect 500 Options

	mean_C...	<new>
495	0.805	
496	0.585	
497	0.61	
498	0.81	
499	0.83	
500	0.79	

3.1002: S: OK. So,
3.1003: looks like it's basically a blob. Um... /cs/ [OK]
3.1004: [00:39:05] It's like a volcano. [/c/] It looks like it's mainly in
3.1005: between -- I'd say, like, point four and
3.1006: point six. There's a little bit of a dip
3.1007: with the point five, but um, I'd say the
3.1008: majority of it is like around that area.
3.1009: [OK]
3.1010: So. [Great.] And it doesn't really have any
3.1011: distinct -- just kind of is like one, like,
3.1012: mountain. [OK] So. [OK]
3.1013:
3.1014: I: So can you draw
3.1015: that?

3.1016:



3.1017: S: Yeah. { /inaudible/ ... four ...}
3.1018:
3.1019: I:
3.1020: [00:39:50] And, uh, compared to, um -- and so how does it look
3.1021: compared to what you saw at N equals 100?
3.1022:
3.1023:
3.1024: S: Um, it's definitely less variable, I would
3.1025: say. [OK] Or, it just like varies less,
3.1026: like there's less of it being like
3.1027: shooting up to like a certain mean or
3.1028: something, and it's just a lot more of
3.1029: like an average, um. I would say like it's -- it
3.1030: kind of just like is a regular about
3.1031: curve, almost.
3.1032:
3.1033: I: Okay, okay.
3.1034:
3.1035: S: So. Instead of having
3.1036: [00:40:22] like multiple ones.
3.1037:
3.1038: I: Okay. Um, and that's what
3.1039: you mean by varying? It doesn't have as
3.1040: many kind of [Yeah] separate peaks? OK. [mh] And, um, now about how
3.1041: likely is it to get point nine or higher?
3.1042:
3.1043:

S: Um, I'd say it's pretty unlikely. [OK] I mean -- I would say it's unlikely. It's not unheard of and it's not like extremely rare, but it's definitely more unlikely than, uh, it was previously. Um, it kind of tapers off like right around [00:40:55] point nine, so. [OK] It could happen, but you're more likely to be more towards like that fifty cent area.

I: OK. And so now -- um, so again when you do the actual study, you'll only be able to collect one of these. So if you did 200 games, do you think that would be enough information, um, to tell whether the claim was accurate or not?

S: Um, I don't think so, just because it still does vary quite a bit, um, -- there's -- I [00:41:26] mean it goes from around, like, point two to point 9ish. [OK] Um, and so that's quite -- just, like, thinking about what the average is for the payout, like each one of those dots is, like, the mean for 200. [mh] And there's 500 of them and it's still pretty like -- there's still quite a bit of a difference between all the means. So I'd say do-- just doing 200, I don't think it would be [00:41:50] enough, um, at least not with like this -- whatever, like, their probabilities and whatnot are, so.

I: Okay. And, um, how many -- how many trials do you think you would need to be able to need to be able to be sort of more sure that this was, um -- to be more confident on the-- whether the machine was accurate?

S: Um, I would

probably say doing maybe like -- maybe 500 would be better? [OK] Um, yeah. I would say maybe like 500, maybe a thousand. [OK] [00:42:31] Something at least doubled what it -- whatever it is, right -- dou-- doubled like two hundred at least, I would say. [OK] Just because, the more the better. /c/

I: Okay, okay. And so, um, um, based on -- so you -- you've -- so now you've sort of seen what happens in kind of two possible machines. [mh] So which one of these -- for which one of these machines was it easier to tell that the claim was false? [00:43:01]

S: Definitely machine 1 because each time, like, even increasing it, like it was still quite a bit lower -- like drastically lower than, um, what this one was. This is mostly like arou-- it was kind of centered around like point 5, while this one most of the results were in between point 1 and point 2 which is not that much, so [mh]...

I: And so if they wanted to -- um, if they wanted to make -- if they designed the machine in [00:43:32] a way that made it kind of really hard to tell whether the claim was accurate or not, how would they design the machine? [Um...] How would -- how could they make it hard to detect whether it's accurate or not?

S: I would probably say like making it more likely to get twenty dollars, because then it'll have a lot more like bumps and everything like that, so you won't really know like how likely it is, like -- yeah, you

[00:43:59] could have an average payout of like say 20 cents, but you could also have an average payout of like $2, and like the --in between that like you're closer to like -- say, like a dollar or something like that. [OK] So it definitely like -- or having even more $1s, um, than anything, cuz then it will kind of shift it over instead of more like the zero like even having it [00:44:21] just, like, a dollar it'll definitely shift more, so. [OK] But.

I: So now you've seen a couple examples of what things look at -- look like at 100 and 200. Um, what do you think -- in general, so not knowing anything about the machine, um, besides that it has 1, 2, and 20, how many -- do you have a kind of final sort of sample size that you recommend going in to do the study?

S: I'd probably say at least having like -- we'll say like five [00:44:51] hundred trials. Because if you don't really know the -- what like the probabilities are, like, it may work -- like, only doing like a hundred may work for like one, but if it's like this one like -- it -- just doing 200 like it's a very very different graph, so I would say probably like doing like four hundred, five hundred, something like that, I feel like would give a much better result so.

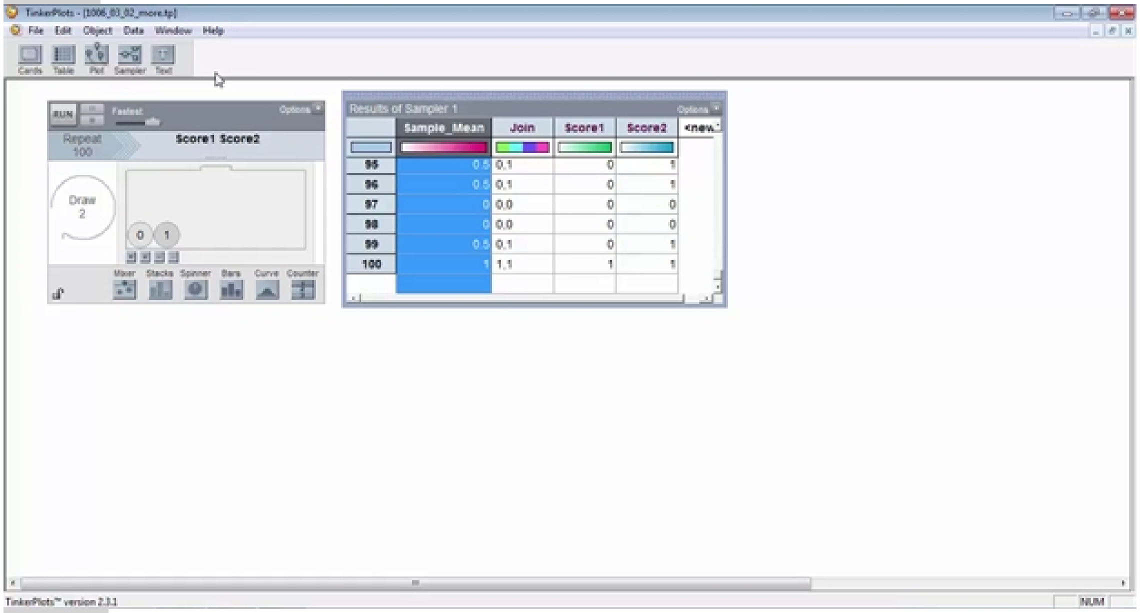
I: Okay. [00:45:16] And, um -- so, um, what if -- um, so let's say I'm -- you know, we've -- I'm at the casino and I decide to be tricky and so I only cheat the customers just a little bit. Um, so I make the average payout just 85 cents. Um, so really close to 90, but not quite 90. Um, if -- if I really had set up the machine like that, how many, um, games do you think you'd need to play in order to detect, um, that the claim was [00:46:03] inaccurate, to detect that I actually made it 85 cents?

S: Um, I would say to like have exactly 85, you would definitely have to do a lot. /c/ Like probably more than like, say like 500, because I feel like 500 would give probably like a good average but for the pr-- probability maybe like adding another 200 trials or like doubling it, um, just because getting it very accurate, like, it could still vary, but I [00:46:33] think -- the more the better, because like you'll just be able to get -- like as you do more, you'll always get, like, to be a more accurate, um, and closer to what it actually is. Uh, so I would say more than 500.

Growing More Means: 0-1

I: Okay. Okay. Okay. So that's all for these casino problems for now. But from now on we're gonna be looking more at, um, these distribution of means, so the next [00:47:08] activity kind of relates to that. So -- but it's set up in a slightly different way than what we've seen before. So, um, {just delete these right now}, so, now I've set -- usually /your statistics class/ you have the Draw set to one [mh], but in order to kind of think more about these distributions of means I've set the draw to two. And basically what happens each time is I'll draw one from there, and

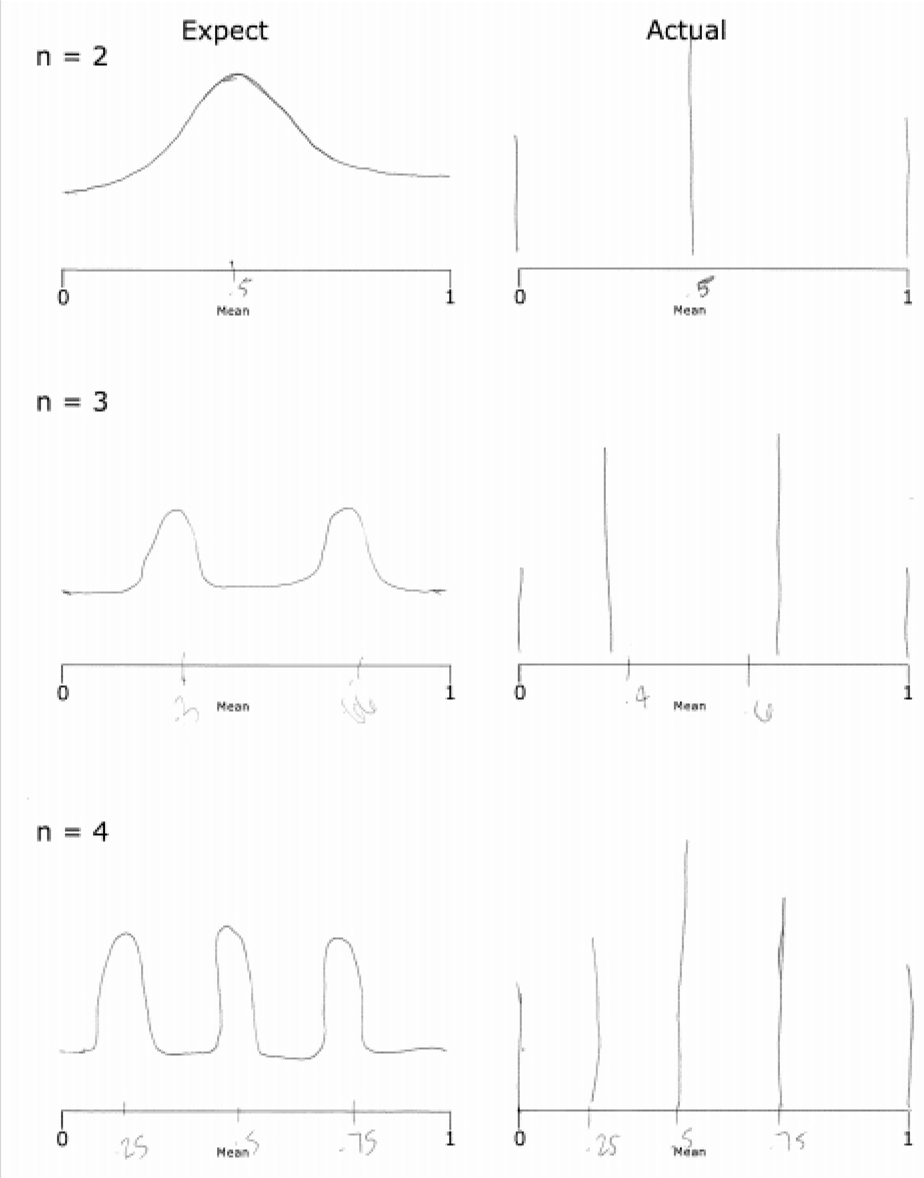
[00:47:40] then I'll draw another one. [OK] Draw one, and then another one. And you can see each time, um, I've calculated the mean of those two, so it's just a -- the mean of a 0 1 variable. Um, so this time I drew a 0 and then a 1, and so that's a mean of 0.5, and you can also see the individual draws here in these columns. Um, and then we'll just keep on doing that, so I'll do that a hundred times. And again you see, like, okay, I drew a zero and [00:48:09] then a zero, zero then a 1, .5, 1, that kind of thing. [mh] Um, so, um, what I'd like for you to do is to draw -- so what do you think this plot of means will be. So if I plot this column, uh, what do you think that plot would look like for N equals two?


TinkerPlots - [1006_03_02_more.tp]
File Edit Object Data Window Help
Cards Table Plot Sampler Text
RUN
Fastest
Options
Repeat 100
Score1 Score2
Draw 2
0
1
Mixer Stacks Spinner Bars Curve Counter
TinkerPlots™ version 2.3.1
NUM

Results of Sampler 1

	Sample_Mean	Join	Score1	Score2	<new
95	0.5	0,1	0	1	
96	0.5	0,1	0	1	
97	0	0,0	0	0	
98	0	0,0	0	0	
99	0.5	0,1	0	1	
100	1	1,1	1	1	

S: Um, I'd say it'll probably be -- like, they'll be some zeros, and not -- but it'll -- I think there's gonna be a lot of, um, -- lot -- like more point 5s than there would be like say a 0 [00:48:50] or a 1, so.



I: Okay. Um, and so why is that?

S: Just because, like, if you -- you could draw a zero and a one, or a one and -- I mean, like a one and -- you could draw a zero and a one, um, just it's more -- I would say... Like it's -- I would think it would be more likely dr-- get, like -- have an average of like point five then say like a one, because you would have to get two ones [OK], [00:49:18] um, instead of like a 0 and a 1, or a 1 and 0. [OK] So.

I: Okay. Um, so we can plot that just to see what happens. And you can run it again just to sort of see what the pattern is.

S: I'd say it varies, but it's pretty much just like 0.5, and then 0 or 1.

I: Okay. Um, so you can draw that.

S: I'll just draw like, the line, cuz it'll be easier to...

I: Okay. And what do you think will happen -- [00:50:11] so I can increase this to three, so that I draw three out, and take the average. What do you think [mh] the average -- what do you think the distribution of means would look like in that case?

S: Um, I'd probably say that one -- like one and zero are gonna be lower. And then I guess it'd be like -- point... point three. Yeah. Maybe. We'll go like that, just for now, I don't [00:50:42] really -- I don't know the exact numbers, but I'll probably -- I would probably say, like, it'll be lower and then it'll go up and then go up again, and then go back down. Like with the two, like, if you, like, like one or a zero, or... maybe there, no, yeah. I'll probably leave it at that.

I: Okay. And so why are point three and point six six more likely than one or zero.

S: I just thought of like, like one out of three trials -- like 0.33 -- my brain isn't really working for math, but I was just thinking, like, [00:51:17] one-third. [OK] So. [OK] Yeah.

I: No, but why -- why would this -- so you have a -- a bump there that's higher than zero. [Yeah] So why -- why would point three three be more likely than zero?

S: Because to get a mean of zero, you'd have to get zero three times, which, like, it'd be a lot easier getting like two ones, or like at least like more than just like two of something, rather than three. Um, it just like -- I just think it would be more [00:51:48] likely instead, um, of getting three zer-- a three zeros. Like, you could get two ones, or two zeros, and then just like one zero. [OK]

I: And so what are -- are these b-- are these the same height?

S: Yeah.

I: Okay. And so why would they be the same height?

S: Um, I think that there's -- it's -- like, you -- the 0 and the 1 is the same probability, like of it being drawn, [OK] so like there [00:52:17] shouldn't be any difference between like the chances of getting two zeros and then also like -- or two ones. Like there shouldn't be any difference between those two.

I: Okay. Okay, so let's -- we can just change it to three, the draw. And then see what happens.

S: So then... it's pretty much what the previous one was -- it just has like another line, and then 0 & 1 are lower than, um, the two [00:52:47] numbers in between.

I: Okay.

S: {oh point four, point six...}

I: Okay. And what do you ex-- what do you, uh, think would happen with N equals four?

S: Um, I think there would be, like, three. So maybe, like, point two five, point five, and point seven five, because that's a fourth. So it'd probably be -- I'd say probably the zero would be low again and then there'd be 25 -- or 25 cents, 50 cents, and 75. Um, and they would all be equal height as well.

I: Okay. And why is that?

S: Um, because I think [00:53:38] there's equal -- you're -- you have the same chance of getting a zero or a one, um, for each time. It just depends on how many times you actually get them.

I: Okay. Great. Um, so

you can change and see it.

S: So point five was higher. So then /inaudible/ point 25, .5, and point 75. So one and zero were lower, and then -- we'll say, like, I think it's to point two five, so I'll put it there, was higher than zero, but 0.5 was higher [00:54:18] than the other one, and then slightly higher too. But I'd say if we did, like, multiple trials it'd probably be -- they would -- k-- like go up and down, but they'd probably be pretty similar to each other.

I: Okay. And why is that?

S: Um, I think it's the same likelihood of getting either or. So.

I: Okay. Um, so -- um, so is that what you expected, what you saw there?

S: Um, I expected them all to be equal, but the point five was h-- much [00:54:52] higher than the rest of them. [OK] So.

I: And, uh, why do you think that would be?

S: Um, I think it's a lot easier getting something in the middle, um, then getting say like l-- it leaning to one side or the other. [OK] Um, just because like if you get, like, two ones and two zeroes then that's gonna be fifty cents, or -- you're -- like it'll be less likely that you'll get like all four ones, so. [OK]

I: And so, um, what would -- what [00:55:27] would happen if we kept going, if we went up to five six seven eight nine ten? What would that look like?

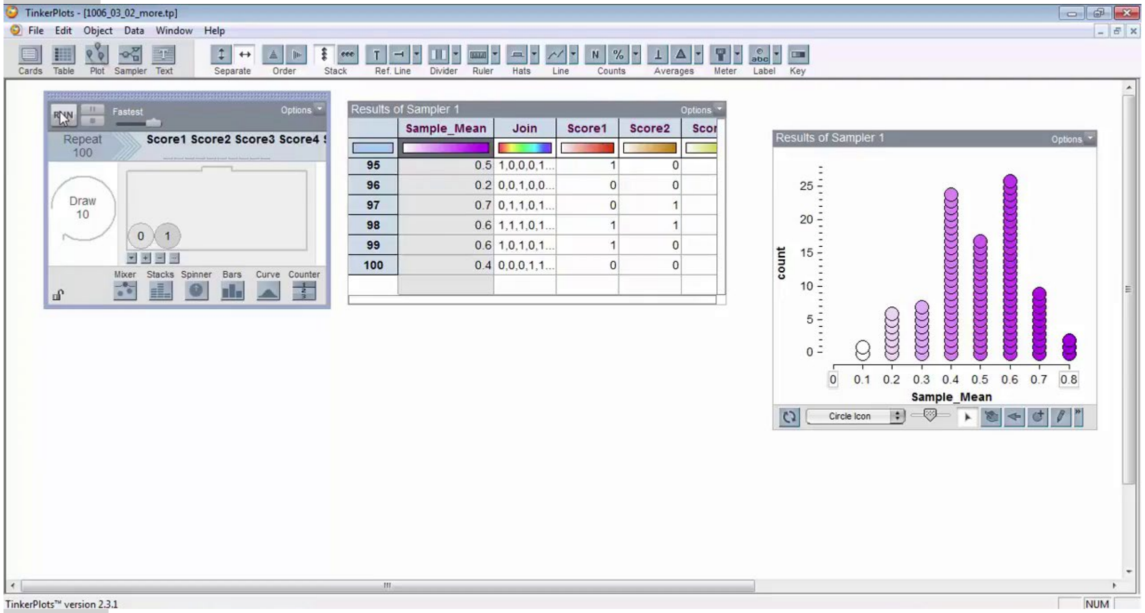
S: Um, I think they would all start centering around, maybe, around fi-- .5? But it definitely -- like there'd be more, um, like, lines each time, and it'll probably have like a center around -- somewhere in the middleish area, I would say. So whatever the fraction will let it be. [mh] So.

I: Um, and so let's say if -- if we put a divider down [00:56:00] between -- so let's not actually do it, [Yeah] but just, uh, if you put a divider down between 0.25 and 0.75 or something like that [mh]. Or point four and point six. Um, um, would happen to the kind of percentage in that divider as the sample size increased?

S: Um, I think it would probably -- it would get smaller, I think. [OK] They would still -- I still think it would be more than like other numbers, but there's a lot -- you're a lot more ch-- [00:56:33] like choices, or like you have a lot more areas that you -- like there's more fractions available, like, for this one it's like -- you only have a fourth of a chance, like 1 out of like -- you can only really have like four -- like 3 -- like I guess, like, I guess 5 different choices, but as you get like to like 8, you'll have 9 different things that you could have, um, [OK] 9 different probabilities, so. [OK]

I: Um, so let's just -- let's [00:57:00] go to a higher number, you can choose anything you want.

3.1444:
3.1445:
3.1446: S: We'll do ten.
3.1447:
3.1448: I: Just to see.
3.1449: And so, um --
3.1450: and so what's the percentage in the -- in
3.1451: the center there about.
3.1452:
3.1453: S: Um...
3.1454:
3.1455: I: Or we can do a
3.1456: divider if you want to.
3.1457:



Results of Sampler 1

	Sample_Mean	Join	Score1	Score2	Scor
95	0.5	1,0,0,0,1...	1	0	
96	0.2	0,0,1,0,0...	0	0	
97	0.7	0,1,1,0,1...	0	1	
98	0.6	1,1,1,0,1...	1	1	
99	0.6	1,0,1,0,1...	1	0	
100	0.4	0,0,0,1,1...	0	0	

3.1458: S: It looks like -- let's see it'd
3.1459: be about like 15ish percent -- at least
3.1460: at point five, because there's about --
3.1461: there's a hundred trials, so. [mh] And before,
3.1462: [00:57:35] this one was at thirty percent, so I'd
3.1463: say it's probably -- it gonna de-- it would
3.1464: decrease, but yeah, I would say-- I would say it decreases,
3.1465: but it definitely -- there's more variation,
3.1466: like -- or there's not more variation, but
3.1467: there's more choices, with, [OK] like smaller
3.1468: percents.

I: And what about the likelihood of zero, or one. How's that changing?

S: Um, I think it decreases. [OK] There's no -- there wasn't even any zeros, um, for this one, so [00:58:11] I would probably say it definitely decreases on the edges. So.

I: And so if we -- what if we, again, thinking about that area between say four point and point six, what if we went up to fifty?

S: Um, I would say that the majority of the answers on the percent would increase for that range, um, just because you'll have less -- less opportunity-- I would say like it'll just kind of get -- the range will get [00:58:43] smaller.

I: You were saying something about opportunity?

S: Like, you're-- it's gonna be a lot less likely that you're gonna get like zero per-- like zero for the mean, because you would have to get like 50 zeros in a row, which is a lot more unlikely than getting like a couple ones and a couple zeros, so.

Growing More Means: 0-1-1

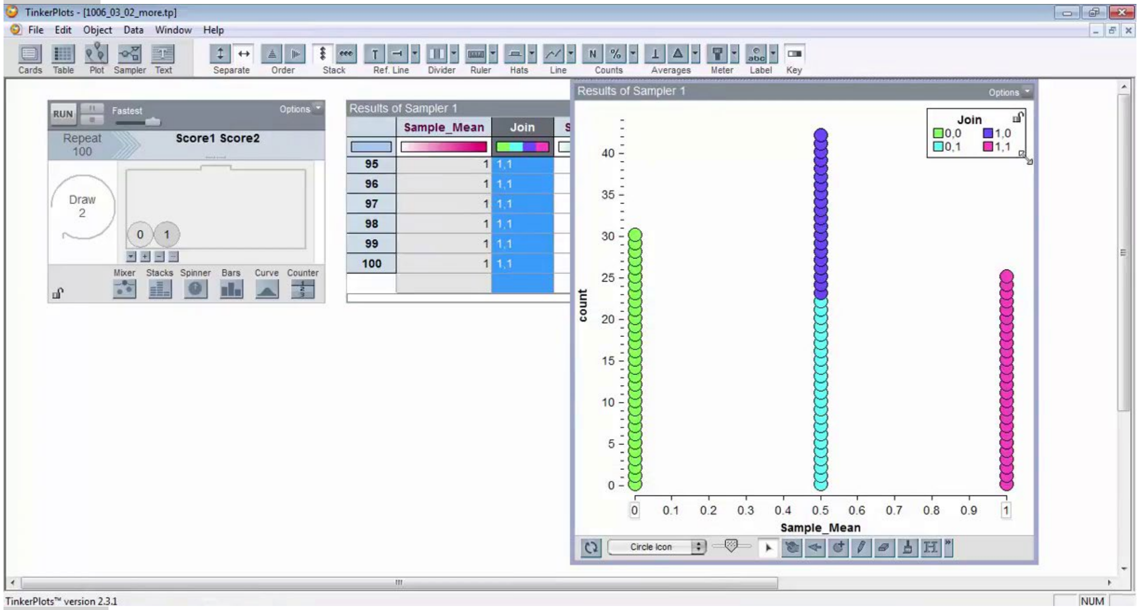
I: Okay. Okay. So we're gonna, um, change this in one [00:59:07] way. [OK] We're gonna add in -- I'm gonna add in another one [OK], but I'm gonna -- I'm just

3.1508: gonna call it one star. But it's still
3.1509: worth one, I'm just using this [OK] to track
3.1510: which one of these we draw. And I will shrink this for now, go back
3.1511: to two. And, um, so again I'm just doing the
3.1512: same thing, draw one and then the other
3.1513: one. Um, and so still if we get one star and
3.1514: zero, that's still point five [OK], and two
3.1515: zeroes is zero, and -- and so forth. Um, so does
3.1516: [00:59:48] that make sense [Yeah] what this is
3.1517: representing, here? [mh] Okay. So, um, I was wondering if you could
3.1518: draw for me, what do you think this would
3.1519: look like, um, at -- what the graph of the
3.1520: means would look like at equals two?
3.1521:
3.1522:
3.1523: S: Um, I would probably say that it'll be more --
3.1524: instead of being more towards like the
3.1525: middle, I'd say it's gonna be more
3.1526: towards like -- I'll say like point 0.75, we'll
3.1527: say, as like an average. So like -- kind of
3.1528: [01:00:23] before how it was -- it'll go up and then go down
3.1529: back to one, but it'll be more over to
3.1530: the -- more towards, um, one just because you
3.1531: have another one in the mix, and so
3.1532: you'll be -- a lot -- it -- you'll be more
3.1533: likely to get one than you would be zero.
3.1534:

3.1535: Growing More Means: 0-1, continued

3.1536: I: Okay.
3.1537: Oh. I'm sorry I actually forgot to do
3.1538: something.
3.1539: So let's back up for a moment for this [OK]. Um, I forgot
3.1540:
3.1541: [01:00:47] to do something here.
3.1542: Um, so I take that one out. Um, here we have this. And I just wanted
3.1543: to show you another way of looking at
3.1544: this. [mh] Um,
3.1545: so we can -- um, I'm just gonna highlight this.
3.1546: Um, well, I will -- um, I'll sort the joins. Um, and then I'll just

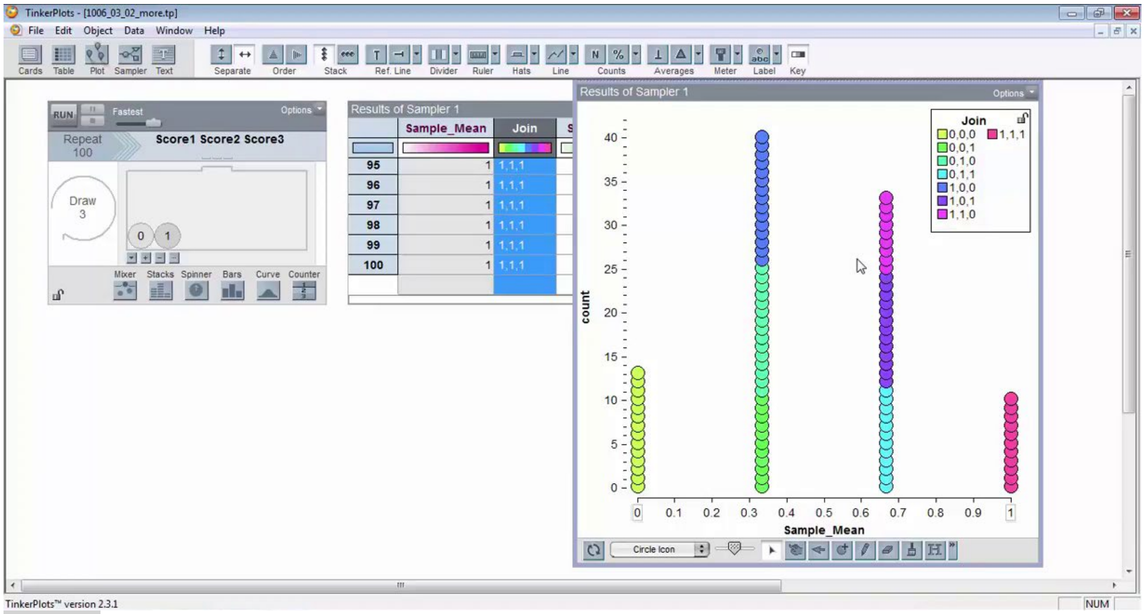
3.1547: show all of the -- so the Join's showing
3.1548: which happens -- the 0 then 0, 0 then 1, 1
3.1549: then 0, or 1 then and 1. And what do you
3.1550: see when I kind of plot things this way?
3.1551:


TinkerPlots - [1006_03_02_more.tp]
File Edit Object Data Window Help
Cards Table Plot Sampler Text Separate Order Stack Ref. Line Divider Ruler Hats Line Counts Averages Meter Label Key
RUN
Fastest
Options
Repeat
100
Score1 Score2
Draw
2
0
1
Mixer Stacks Spinner Bars Curve Counter
Results of Sampler 1
Options
Join
0,0
1,0
0,1
1,1
count
40
35
30
25
20
15
10
5
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Sample_Mean
Circle Icon
TinkerPlots™ version 2.3.1
NUM

Results of Sampler 1

	Sample_Mean	Join
95	1	1,1
96	1	1,1
97	1	1,1
98	1	1,1
99	1	1,1
100	1	1,1

3.1552: S: Um,
3.1553: [01:01:36] that it's a lot more likely that you'll
3.1554: get either a 0 1, 1 0, like, because
3.1555: that'll equal the same mean, than you
3.1556: would for a 0 0, 1 1.
3.1557:
3.1558: I: Okay. And if I go up
3.1559: to N equals 3.
3.1560: Again, I'll sort this.
3.1561: And what do you see here?
3.1562:


TinkerPlots - [1006_03_02_more.tp]
File Edit Object Data Window Help
Cards Table Plot Sampler Text
Separate Order Stack Ref. Line Divider Ruler Hats Line Counts Averages Meter Label Key
RUN
Fastest
Options
Repeat 100
Score1 Score2 Score3
Draw 3
0 1
Mixer Stacks Spinner Bars Curve Counter
Results of Sampler 1
Results of Sampler 1
Options
Join
0,0,0
0,0,1
0,1,0
0,1,1
1,0,0
1,0,1
1,1,0
1,1,1
count
0
5
10
15
20
25
30
35
40
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Sample_Mean
Circle Icon
TinkerPlots™ version 2.3.1
NUM

	Sample_Mean	Join
95	1	1,1,1
96	1	1,1,1
97	1	1,1,1
98	1	1,1,1
99	1	1,1,1
100	1	1,1,1

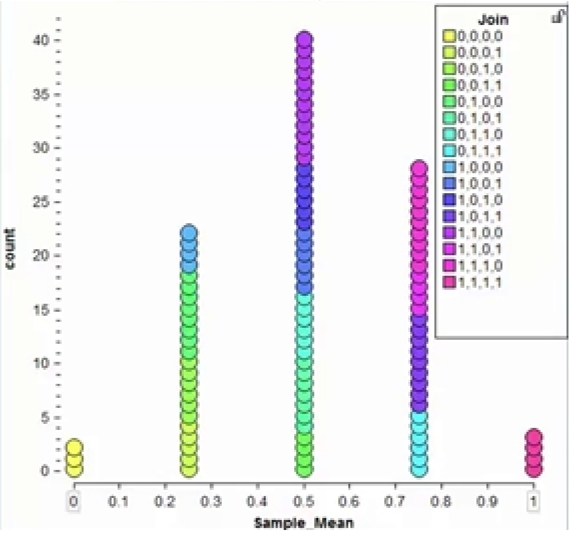
S: Um, I'd say, it's a lot more likely to get a combination, um, of like the mean equaling 0.25 or point, um, like or 0.33 or 0.66. It just [01:02:24] just like different combinations, and they all I think have, like -- they look about like they're equally likely to like get that comb-- combination, so like say like, a 0 1 0, or a 0 0 1, like, they get the same mean, but, um, you're equally likely as -- to get one of those, so.

I: Okay. So just say a little bit more about that. You're equally likely to get, um...

S: Like a -- like a mean equaling like 0.33, from like 3 different [01:02:50], um, like things. Like 3 different numbers. So. [OK] Or like number of combinations, I should think.

I: Okay. So are you saying that each of these combinations -- 0 0 0 -- at -- each one of these combinations is

3.1588: equally likely? Or...
3.1589:
3.1590: S: For, um, just like 0.33, and
3.1591: 0.66, like. I'd say it's probably
3.1592: pretty close for like 1 -- getting a mean
3.1593: of one and zero, but I think it's more -- I guess all
3.1594: of them would probably be equal, but
3.1595: [01:03:22] getting the mean to equal those two, it's
3.1596: more likely to get [OK] in between.
3.1597:
3.1598: I: And you
3.1599: said -- again, you said all of them are
3.1600: probably equal? All of --
3.1601:
3.1602: S: Yeah. For, like -- at least
3.1603: like -- they all have an equally likely
3.1604: chance of being drawn, but the mean -- it's --
3.1605: you're more likely to get the point 3 3
3.1606: or point 6 6, [OK] instead of like 1 or 0. [OK]
3.1607:
3.1608: I: We'll just look at n = 4
3.1609: this way as well. Again,
3.1610: sort.
3.1611: [01:04:01] And what do you see here?



3.1612:
3.1613: S: All of like, the -- little
3.1614: stacks are kind of equal. The only one
3.1615: that's like much larger is, um, like the -- the

purple one for -- so that would be like the 1 1 0 0, that's the most likely. At least in this trial it was. Um, but they all are about the same, like, number. [OK] So.

I: And how do the number of combinations in each stack compare?

S: They're about the same. They're just -- it's just more likely to a mean [01:04:36] of 0.5?

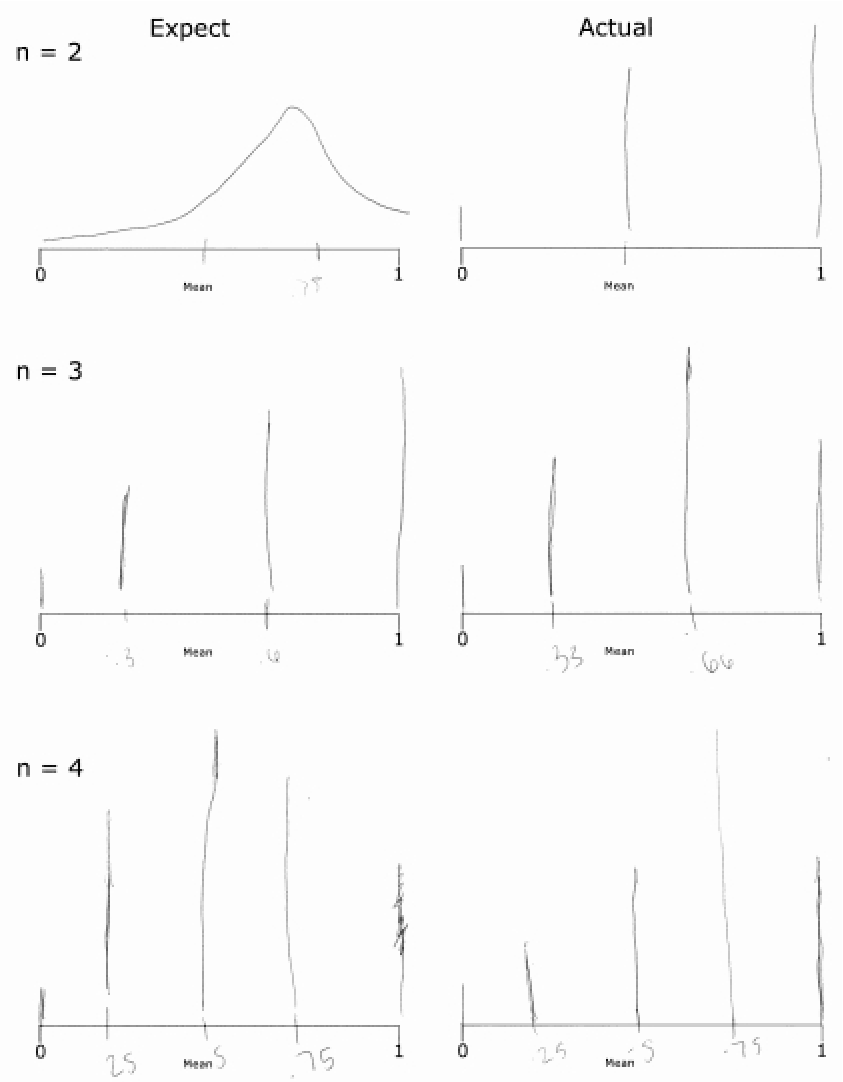
I: And why is that?

S: Because there's more combinations that you can do to get, um, a mean of 0.5.

Growing More Means: 0-1-1, continued

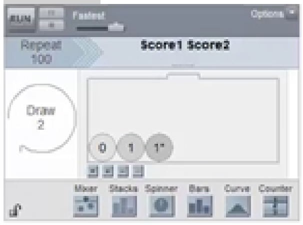
I: Okay. Okay. OK, so we're gonna /c/ [Yeah] =now we're gonna back to--=

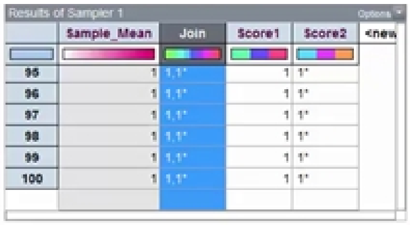
S: =No -- no, we're good.= /c/

3.1638: 

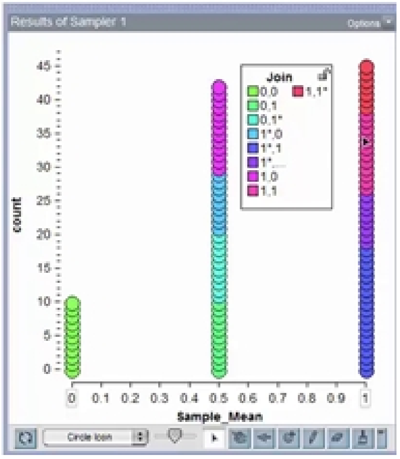
3.1639: I: -- go back to the
3.1640: other one. Um, so, um, let me add back in my 1
3.1641: star. Um, and we can -- um,
3.1642: we can now -- um, so yeah. Let's see what
3.1643: happens. So you can change the Draw back
3.1644: to two. If you want to see it the same as
3.1645: the other way, you can sort by the join. [/inaudible/]
3.1646: So go -- click on join, [Eh --] and then [Rows --] Sort Rows Ascending. [OK] Yup.

3.1647:




Results of Sampler 1

	Sample_Mean	Join	Score1	Score2	<new
95	1	1,1*	1	1*	
96	1	1,1*	1	1*	
97	1	1,1*	1	1*	
98	1	1,1*	1	1*	
99	1	1,1*	1	1*	
100	1	1,1*	1	1*	



3.1648: S: {Make this one smaller} /S shrinks key/ [/c/] There we go.
3.1649: [01:05:30] So this one was -- a lot -- it's more likely
3.1650: to get point .5. or .1, but .1 is definitely more
3.1651: common. Just slightly, though. So I would
3.1652: probably say... that... so actual would be this. So this is still at 0.5,
3.1653: and point 1, which I'd say having like
3.1654: point 1 -- I'm more surprised about
3.1655: the point 5 just because, like, there is a
3.1656: lot more ones in the mix. Um, and less zer--
3.1657: and less zeros, so [OK] but, um, it makes sense that
3.1658: it's definitely skewed more on like
3.1659: [01:06:16] towards 1 than it is zero, so.
3.1660:
3.1661: I: Okay. What --
3.1662: you can -- or -- look at different numbers of
3.1663: trials if you want.
3.1664:
3.1665: S: OK, so then... do you mean =by changing the, like --=
3.1666:
3.1667:
3.1668: I: =Or, like just, just= -- you can just click Run again, [All right]
3.1669: just to see what it looks like.
3.1670:
3.1671: S: Yeah, so
3.1672: it kind of goes back and forth. Like one's -- so I'd
3.1673: say they're probably about, like, equally likely,
3.1674: as getting, like, point five as the mean, or

one as the mean, so.

I: So one thing that'll help is -- so, if you sort by Join again. [OK] Um, we [01:06:46] again do the borderless icons. We can kind of see the colors a little bit more clearly. Gosh. Are there two -- [Yeah] so actually, like, I have a color deficit. OK.

S: No you're good. There -- yeah, there's two right there.

I: There's two colors right [mh] there. Okay. I -- =it looks like the same to me.= /c/

S: =No, it's like, it's rea--= it looks -- it's really close. /cs/

I: Okay. Gotcha. All right and so what do you think would happen at N equals three?

S: Um, I would probably say that [01:07:24] maybe this time it'll be closer to -- let's see I think it'll kind of, like, probably be even less likely to get a zero as the mean, um, and then maybe one would be pretty high again, and then, maybe as like point s-- point six would be pretty high -- not as high-- maybe not as high, or, like, about even with like one. And then, yeah. And then, like, maybe 0.3, not as likely but still there. /inaudible/, maybe?

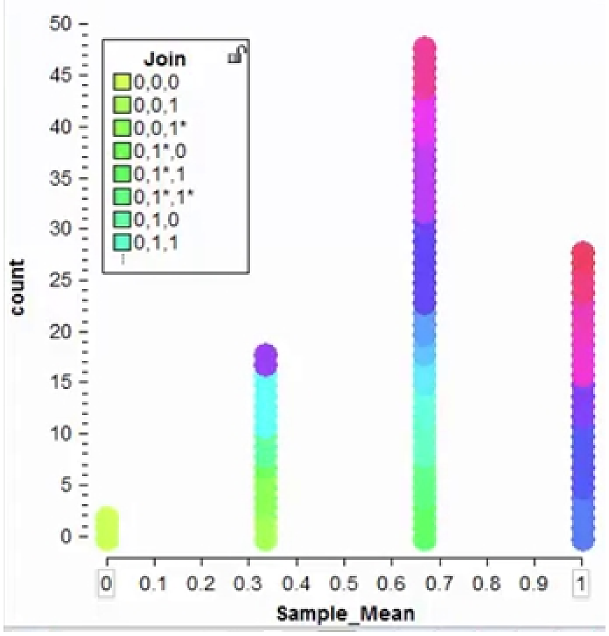
I: OK, and [01:08:12] why is that?

S: Um, just because there's -- you have -- you have three times that you're taking it, and you're gonna be more likely to get a one than a zero.

I: OK. OK, great. So let's see what happens.

S: {Move over here. There we go.}

I: And you can sort, if you want.



S: I'm gonna do a couple more. OK, so it looks like in general that zero is definitely gonna be the lowest, and then 0.3 and point six... six, we'll say. So point three and [01:09:03] point one seem to have about, like, the same likelihood of happening. But it seems like it -- point s-- like the point -- around like the point six that's gonna be the most likely, um, mean that would occur.

I: Okay. And why do you think that might be?

S: Um, cuz it's more -- there's more ones in the mix so while -- before there's probably like an equally likely chance, like, [01:09:30] you're gonna be more likely to have the combination of having, like -- maybe like two ones and a 0 rather than like all ones, or like, all, like, basically all zeros and then like one one. [OK] So.

I: Okay. And what do you think would happen at N equals four?

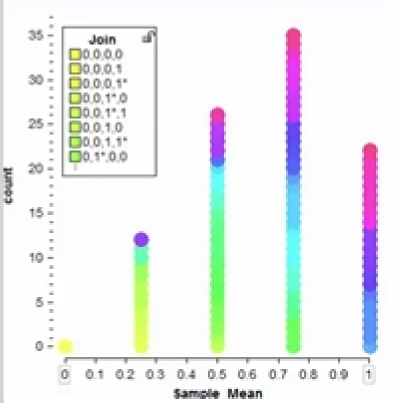
S: Um, I would say probably 0 is gonna be a lot lower, and then I would say, like, it'd be about like point 25, point 5, and then point 7 5. Um, and then probably [01:10:08] -- maybe I think point 5 and point 7 5 will be the most likely and then... then 0.25 and point one would probably be less likely.

I: Okay. And why are you thinking that?

S: Um, just like combinations. It was a lot easier getting, like, a -- maybe like the -- like I guess the point -- I think the point 5 would be higher. Maybe this would be like more equal to like this 7 5. Um, you're gonna have a lot more [01:10:43] combinations getting the point 5, um, so that's where I think point 5 would be the most, and then point 7 5 and point 2 5 are more likely than getting all zeros or all ones for the -- the trial. [OK] I'd probably make this, like, shorter.

I: OK. /c/ Okay. So

let's see what happens.



S: So I'd say it varies, but in general it looks
like 0.75 is gonna be the most common
one. So point one, it looks like there's -- it's gonna be
[01:11:27] likely, but it's gonna be more likely than
0.25, but not as likely as 0.5, and then
point 7 5 is going to be the most.

I: Okay. And what do you think would happen
if we kept going?

S: Um, I'd say it probably -- whatever fraction is,
like, closer to one is gonna always be
like the top. Um, it won't be like centered
in the middle. It'll always be kind of
skewed to the right, but not all the way
[01:12:08] to the right [OK], um, to the highest number.

I: OK. And, um, so if we put a divider -- sort
of say somewhere around this area, uh, what
do you think about the percentage that
would fall in that divider as we got the --
increased the sample size?

S: I'd say it'd
probably have like the majority, so like, /yawns/ maybe

3.1812: /around?/ like 50% or 75%. Maybe like
3.1813: something 75 -- 75%. Like if it's the
3.1814: majority of it, so.
3.1815:
3.1816: I: Okay. And it -- as we kept
3.1817: [01:12:41] on increasing with that percentage stay at
3.1818: 75%, or would it change from that?
3.1819:
3.1820: S: Um, I
3.1821: probably say it's still probably gonna
3.1822: be around the same... [OK]
3.1823: Yeah. Probably around the same maybe
3.1824: decrease a little bit, just because it's
3.1825: like -- it'll get more chances to get those
3.1826: two, but it's definitely gonna be the
3.1827: majority. /c/
3.1828:
3.1829: I: Okay. Okay. Sounds good. So
3.1830: that's actually all for today. [OK]

Appendix D4: Growing Possibilities and Many Means

Introduction

4.1:
4.2: S: I know how to use Legos.
4.3:
4.4: I: /c/ Excellent.
4.5:
4.6: I: I've done it
4.7: [00:00:03] many times. /c/
4.8: You ready to start?
4.9:
4.10: S: Yup.

4.11: Growing Possibilities: 0-1 Building Blocks

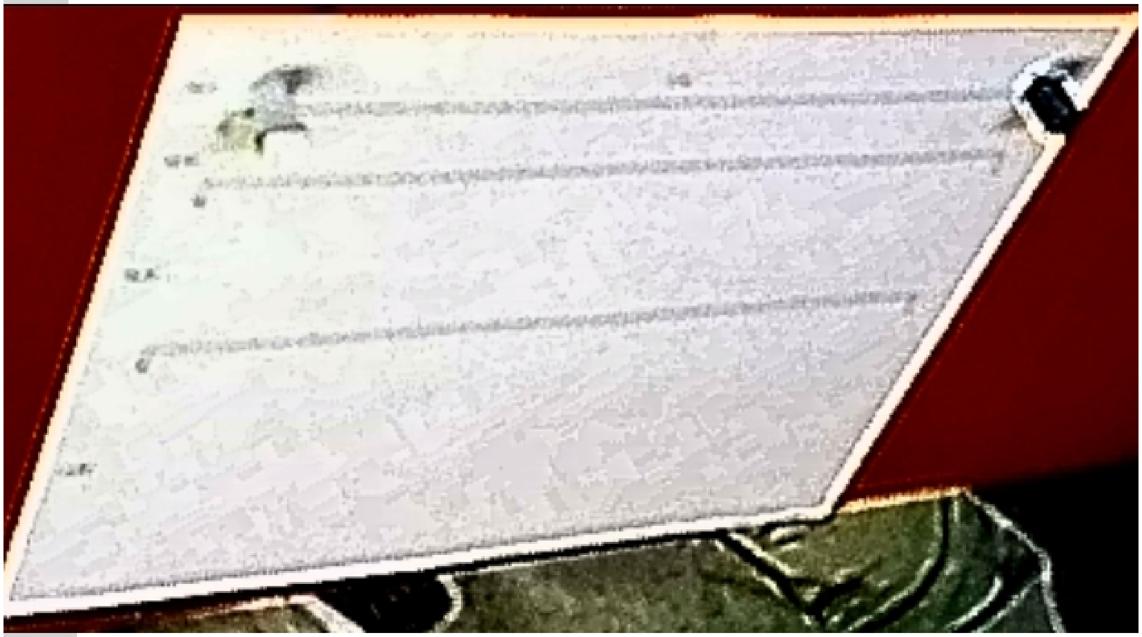
4.12: I: Cool. {I'll actually move this up a little bit. OK. /inaudible/}
4.13: Okay,
4.14: so we're gonna think again about the
4.15: situation where we're drawing

blocks from the box. So here I have, uh, one white block, and one black block in the box. So just like before I would shake the box, randomly pick the block, and put that block back in the box. So here again the [00:00:43] black blocks are scored as a one and the white blocks are scored as a zero. Um, again we're interested in the average score at different sample sizes. So now we're gonna to create some graphs of the different possible outcomes, and the means of those outcomes. So here at N equals one, um, what are the possible outcomes that we can get in the box?

S: Um, either a black or then -- or a white block.

I: Okay and what are the [00:01:18] means of each of those possible outcomes?

S: For just like grabbing it once -- like with -- with that -- like point fi-- or it'll either one or zero [OK] for the first round.

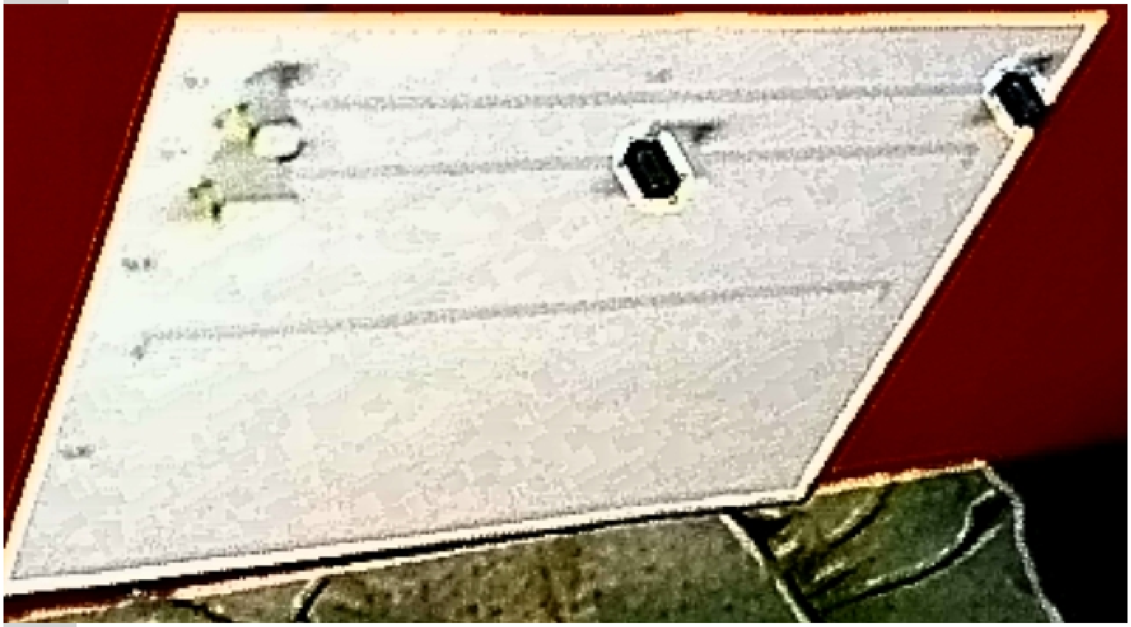


I: Okay. So, um, we'll place a white block at zero, and a black block at one. Like this. Um, so now we want to look at all the possible outcomes when we draw two blocks from the block. Again, drawing them [mh] them [00:01:51] one at a time. We can represent the different outcomes by sticking the blocks together. Um, so if we draw a black block, and then a white block, we can stick them together like this. {Black block then a white block.} What would the mean be if we drew one black block -- a one -- and then a white block -- a 0?

S: Point 5.

4.59:

4.60: I: Um, so we place this block here. We'll
4.61: say that's point five. Um, so now I'd like
4.62: you to put together blocks to represent
4.63: [00:02:33] drawing two -- um, two white blocks in a row.
4.64: Position those on the paper in the same
4.65: way. Um, are these all of the possibilities
4.66: that end in a white block?



S: Does it matter, like, the order it's drawn? Or are we just looking at, like, the number?

I: So I'm looking at the ones that end in a white block.

S: Uh-- that'd be it, yeah.

I: Okay. So, um, how these, um -- these, uh, possibilities the ones that end in a white block relate to the possible [00:03:10] outcomes at N equals one.

S: Um, they have a white block. /c/ Um -- sorry, can you repeat the question?

I: Sure. So we have two, uh, possible outcomes that you have here at N equals 1, and then here at N equals 2 we -- we have all the options that end in a white block. So how do these relate to these up here.

S: There was only one way where you could get a white block -- like it ending in white block for N equals 1. With N equals 2 you [00:03:47] have 2 different options that you could potentially draw and how [OK], um -- it take -- get -- getting the white block the second time.

I: Okay. And so if we draw a white block the first time, um, where does it move when we draw our second white block.

S: Um, it won't move. [OK] If it's, like, all white.

I: Okay. And if we start out with drawing a black block, um, where does it move or not move [00:04:18] once we add on a white block?

S: It'll go to the left.

I: Okay. Um, so now please put together all the possibilities that end in a black block and position them on the paper. So again, um, how do these possibilities ending in a black block relate to the options that we had at N equals one?

S: You only had one, um, option of getting a black block, um, with N equals one, and then with N equals two you have two different [00:04:55] options for ending with a black block.

I: Okay. And, um, where -- um, so if we start with a white block, um, where does the sample mean go when we add on a black block?

S: It'll go to the right. More towards one.

I: Okay. And, um, what about when we add a black block on to -- where were -- the -- on to the one?

S: It'll stay closer to one.

I: Okay. So let's stack up the possibilities, just like TinkerPlots.

And, um, what do you notice so far?

S: That [00:05:34] the options are increasing. Or, like, the different, um, combinations you can get.

I: OK. And, uh, which means -- uh, which means would be kind of most likely for drawing two blocks?

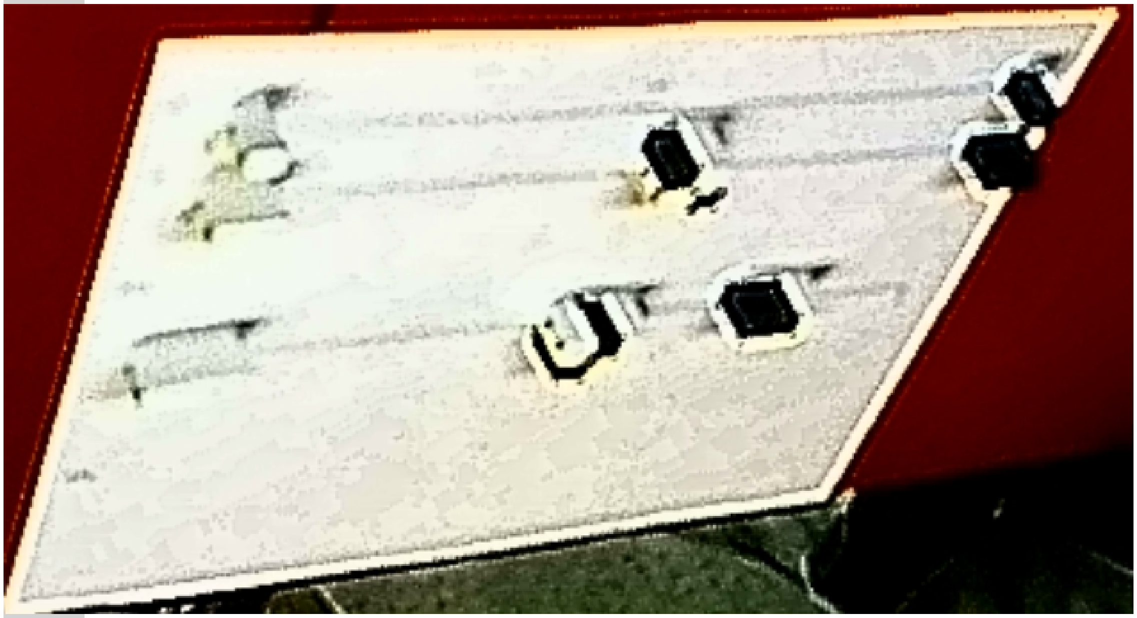
S: Point five.

I: Okay. And why is that?

S: Um, there's two different combinations that you could get to get point five, while, um, to get zero or one there's only one combination that you can get.

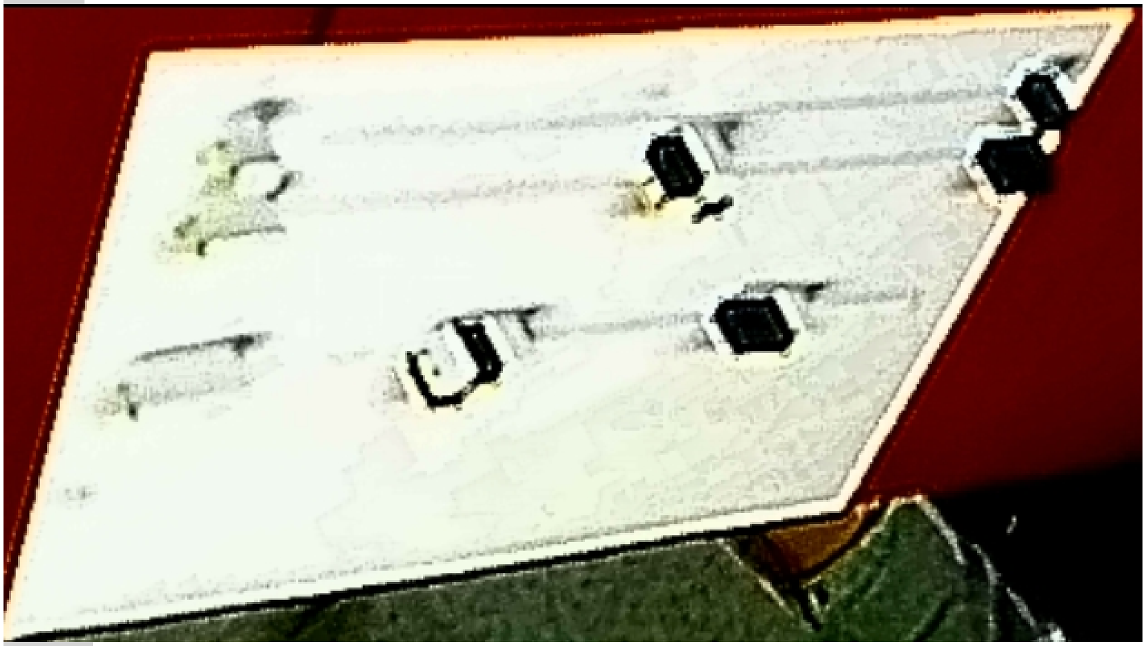
I: Okay. So let's do the same thing for N equals [00:06:00] three. So create all the outcomes that

4.172: end in a white block and position them on -- on the
4.173: paper.
4.174:
4.175:
4.176: S: {Hmm... =think this is--=}
4.177:
4.178: I: =What are you= thinking right now?
4.179:
4.180: S: I'm just
4.181: trying to think of all the combinations
4.182: that I could get. [OK] But I think I have
4.183: them. [OK] Could be completely wrong but, /c/. [OK]
4.184:



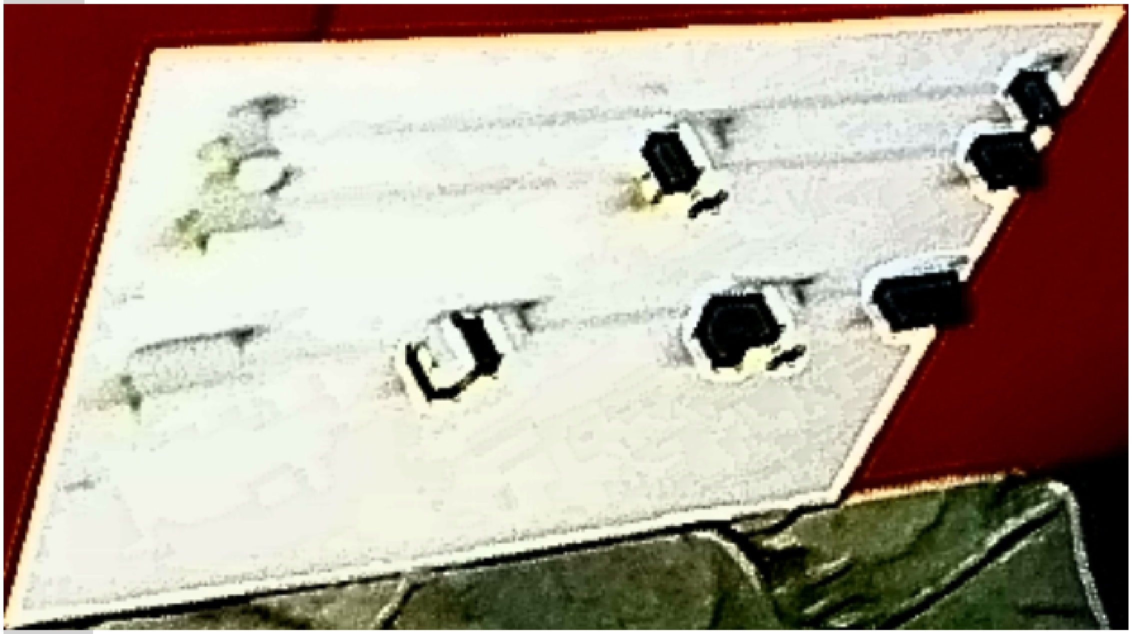
4.185:
4.186: I: And, uh, where -- where are each of these
4.187: placed?
4.188:
4.189: S: Um, so this one's at -- uh, I guess it would be -- these ones wou-- this one would be over here. OK. Um, so this is zero,
4.190: and then this is around like point three,
4.191: [00:07:11] and this is around point six, and then that'd be one.

4.192:



4.193: I: OK.
4.194: And again, um, for each of the possible
4.195: outcomes we saw up here, what happens when
4.196: we add a -- a white block on to them?
4.197:
4.198: S: It'll
4.199: shift more towards zero if you add more
4.200: white.
4.201:
4.202: I: Okay. Um, and let's do -- so let's put
4.203: together the possibilities that end in a black
4.204: block.
4.205:
4.206: S: {And... let's see.}
4.207:
4.208:
4.209: I: So what are you thinking right now?
4.210:
4.211: S: I'm
4.212: just thinking about all of the different
4.213: [00:08:04] combinations. [OK] I thought I had one more in
4.214: my head, but... /c/ [OK]

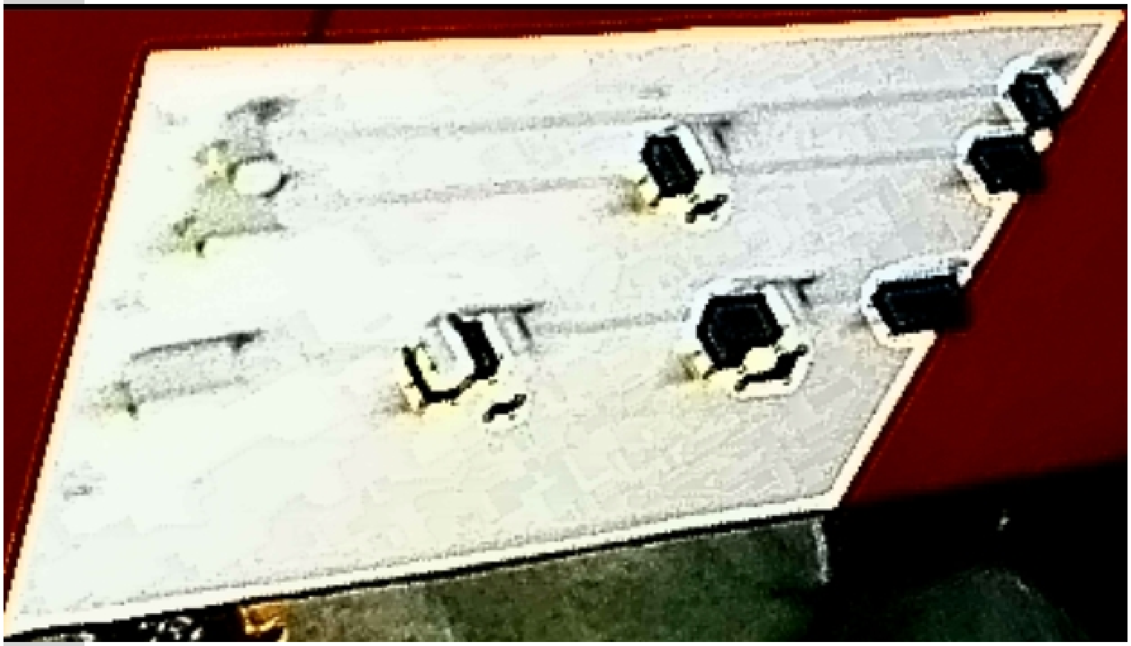
4.215: I think that's it. {Ch -- ch... let's see.} Hmm, yeah, I think that's all of the options. Yeah.
4.216: I'll just go with it for now. /cs/
4.217:



4.218: I: Okay. Um, so again
4.219: for each of the options up here, um, what
4.220: happens when you add on to a black block?
4.221:
4.222:
4.223: S: Um, oh, I -- wait did I... I have one more to put on there. [OK]

4.224:

4.225:
4.226: I: And so what did you -- um, what helped you
4.227: notice that additional block? [Um--] That additional one?
4.228:
4.229: S: I
4.230: was thinking about,
4.231: [00:09:08] um... {let's see. Now I have to see. And then I have to add one more on this side, too. OK.} Um, I was just thinking about how the --
4.232: like, just because the proportions are
4.233: the same, like, you can still get a
4.234: different combination. So I was just
4.235: trying to, like, think in my head. I felt
4.236: like there were supposed to be more than
4.237: just like adding one more, um, [OK] so I was
4.238: just trying to figure out what was -- what would I --
4.239: what was -- what I was missing, so.

4.240:

4.241:
4.242: I: Okay. Okay. So we're looking at the
4.243: [00:09:36] different combinations and seeing what
4.244: mixture would sort of -- um, that you were missing [mh] there for it. Okay. Um. Okay. So I'll
4.245: ask again. So how -- how do each of, um, the
4.246: combinations at n = 2, how did they
4.247: correspond to what you're seeing at N
4.248: equals three when you add on a black
4.249: block?
4.250:
4.251: S: Yeah. Um, so there's more combinations
4.252: that you can get, because, um, even though the
4.253: proportions, like, can, they can vary with
4.254: it, um, you can still get different, like, you
4.255: [00:10:14] could get a black block first and then
4.256: two white, or you can also get two white
4.257: and then a black block, [mh] but, um --and it won't
4.258: affect the mean differently. [OK] It'll be
4.259: the same number for that mean, but it'll
4.260: affect, like, what order it is, and it'll
4.261: become -- it'll make it more like probable
4.262: that it could happen

4.263: rather than leaving all three.
4.264:
4.265: I: Okay. Okay. Um,
4.266: so finally, let's look at N equals four. And so
4.267: [00:10:42] create all the outcomes that end in a
4.268: white block and position them on the paper.
4.269: So can you talk a little bit more about
4.270: what you're thinking as you're doing this?
4.271:
4.272: S: Yeah.
4.273: I'm kind of looking at the ones before,
4.274: and then I'm also thinking of, like, if I
4.275: take one block away, then I could make it,
4.276: like, black or white, or just, like, -- w-- like --
4.277: just like switching them around and then
4.278: also like thinking what could I start
4.279: [00:11:10] with that still ends with like a white
4.280: block, [OK] or what could end with that
4.281: still ends with a -- a white block. [OK] {Do it like this.} Just leave that in the middle and then organize them all later.
4.282: I think that's all of them.
4.283: [OK] So then...

I: Okay. And so how do, um -- what's the correspondence between, um, the outcomes up here, um, when you start a sample like this and then you add on a white block, what happens?

S: Um, you get a lot more, like, you [00:13:20] get mo-- it -- like more, uh, combinations that you could make to give, uh, for, like each, like, section of it. [OK] So.

I: But just focusing on the ones that end in a white block, is there the same number that end in a white block is there are overall here, or is there a different number?

S: There's more.

I: Okay.

S: Then -- that end in a white block with the N equals four.

I: Okay. Um, [00:13:53] are there more that end in a white block than there are total at N equals three, or they're the same or is --

S: There's more.

I: Okay. Okay. Um, and what happens to each of these if you add on to a white block at the end, how does it move?

S: For...

I: So if I started -- so I had one that started white white black, [mh] what happens if I add a white on to it and make it white white black white?

S: It [00:14:31] will be -- I'm not sure what the question is. Because it'll just end in white, and it'll make [OK] the white proportion higher.

I: Okay. Yeah, that's all I was asking. [OK] Okay. Um, and then -- so now put together the possibilities that end in a black block and [OK] position them on the paper.

S: It's -- oops! /S drops one on the floor/

I: Eh, don't worry about it. [OK /c/] We've got lots of black blocks.

4.347:
4.348: S: {I think that's all of them.} [OK] {Let's see.}
4.349: I think that's all of them.
4.350:

4.351: I: Okay. Um,
4.352: so again how do the ones ending in a
4.353: [00:17:25] black block correspond to the ones that
4.354: we see at N equals three? So what happens
4.355: when we add on a black block to, um, one of
4.356: the ones in N equals three?
4.357:
4.358: S: There's more
4.359: options, um, [OK] just with like the
4.360: different combinations.
4.361:
4.362: I: Okay. Um,
4.363: and what else do you notice at n
4.364: equals four?
4.365:
4.366: S: Um, that the three different
4.367: options are all equal
4.368:

I: Okay. And which means would be most likely if I drew four from [00:18:02] the box?

S: Um... it'll be -- you'll be equally likely to either get, like, this proportion, this proportion, or this one. So either like 25% black, or, um, 50% black, or 75% black /from the box, grabbed?/

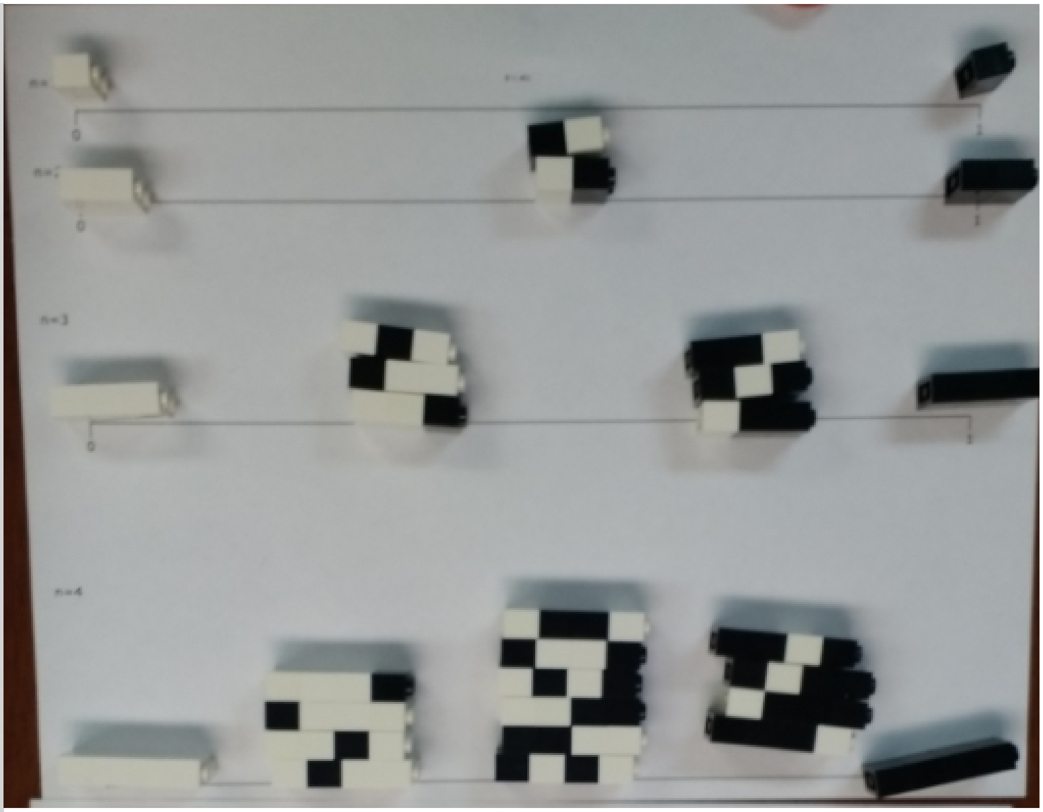
I: Okay. So I'd like to share with you a, um, a -- one strategy for doing this. We're gonna do a little bit more of this, so I just wanted to share another strategy. And you can also [00:18:39] use it to check what you have here. So, if you --

S: Oh, I know another combination.

I: Okay. You can put it. /c/

S: Sorry.

I: Okay. So I guess I'll ask that question. /cs/ Um, which -- uh, which means now do you think are most likely?



S: The -- getting a combination, um, like, 50/50.

I: And so why do you think that is?

S: Um, there's dif-- there's more combinations of getting it fifty-fifty than there is for 25 percent or 75.

I: Okay. And, um, what would you expect to [00:19:40] happen if we kept increasing the sample size?

S: Um, that probably -- the 50-50 chance will -- or, like the 50-50 for a black and white blocks will be higher than, um, like for the even amount of blocks that you have.

I: So if we went up to N equals 10, would you think that zero would be more likely here than at N equals four, less likely, or about the same?

S: Um, less likely.

I: Okay. And why is that?

S: [00:20:14] Cuz there's more combinations an-- there'll be more combinations in the middle that you could get that aren't zero, [OK] so it just put it down even lower for percentage-wise.

I: Okay. And, um, if we sort of again drew a sort of a divider around the center, um, from say 0.4 to .6, or 0.25 to 0.75, um, what percentage of the outcomes might you expect to be near the center at N equals 10? [00:20:44]

S: I would say that for N equals 10, the majority of them will be like -- like within that range. [OK] Um, yeah. I'd say that the majority of them would definitely be in that range. [OK] So.

I: And what about at N equals 50?

S: Um, I think... that... they'll even -- it'll be, um -- like most of them will also still be in that range. I

[00:21:22] think it'll be a higher percentage [OK] than t-- n equals 10, just because the fractions are getting smaller, um. [OK] So, it'll just like put them closer together and t-- closer or they'll be more -- like different of outcomes, like an increase in -- and then in that, um, range.

I: Okay. So again, um, now I'll show you sort of a way of, um, kind of checking what you've done already or doing again in the future. [00:21:51] So if you take all the outcomes at N equals one, um, we could say, oh, which of those outcomes, um -- we want to see all the ones that end in white first. So we take this and we add a white on to it, and then we take this and we add a white on to it and that gives us the two outcome-- that gives us the outcomes like that, and then if we want to see all the ones that end in black we take this one [00:22:21] and add a black on to it and then we take this one and add a black on to it. Does that -- do you have any questions about that strategy?

S: Uh-uh.

I: Um, so could you just talk through sort of checking to see if you have all the combinations by doing the same thing from N equals 2 to N equals 3 to see what happens when you add a white on to each of these and where it goes and what it corresponds to?

S: Yeah. So if you add a [00:22:50] white on to this one it will give you that one. [mh] And then if you add a black on to this one, you'll get that. Um, and then if you add a black on to this one, you'll get { /inaudible/ this one?} Oh, yeah, you'll -- no. I guess I --

I: So you're pointing to this one. Okay.

S: Yeah. I don't think I have =that.=

I: = So that would --= what would that be. Just tell me -- what would the colors be. It'd be black white...

S: Oh, it would be this one. Yeah. /c/ [OK] [00:23:25] Yeah. If I add a black -- or is it me adding a white or a black, now I can't remember. If -- okay. So if I had a white to that -- it'll that. If I add a white to this one, it'll be that one. /c/ [OK] And then if I add a white to this one, it will give me that one. And then if I add a white to this one, it'll give me that one. [OK] Um, then if I add a black to this one, it'll give me that. If I add a, um, [00:23:54] let's see -- if I add a black to this one, then it will give me this one. And then if I add a black to this one it will give me that one, um, and then if I add a black to this one, it'll will give me that one.

I: Okay. And can you do the same for N equals 3 to N equals four?

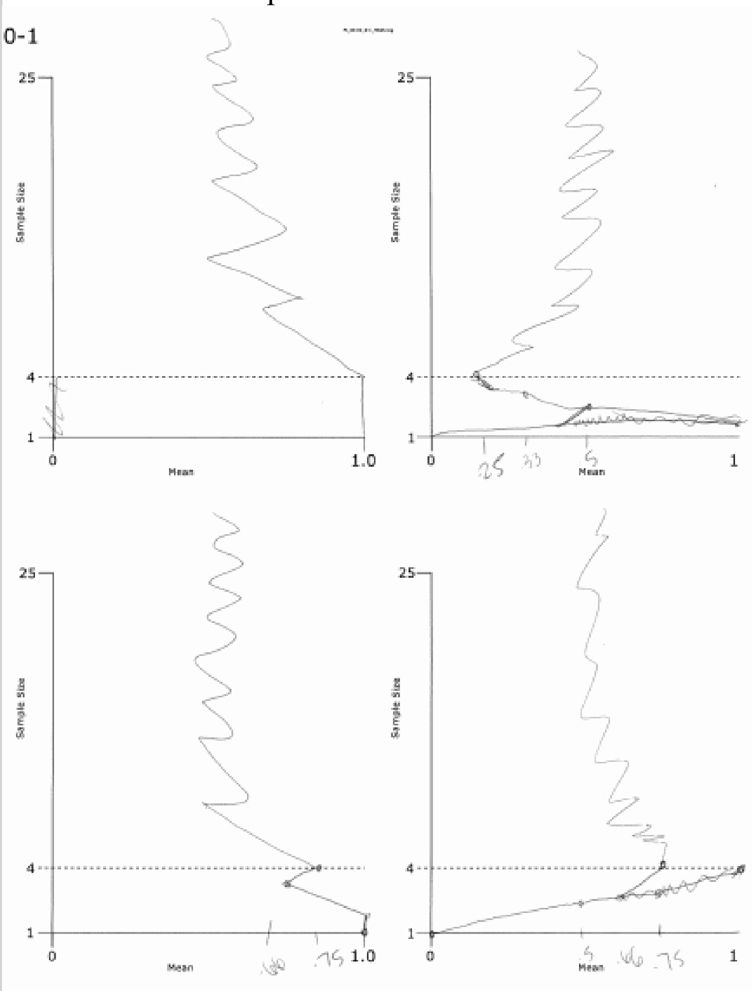
S: Yeah. So if I add a white to this one, it'll give me that one. If I add a white to this one it will give me -- let's see -- two whites. Um, it'll give me that one. [mh] And then [00:24:32] if I add a white to this one, it will give me, um, this one. And if I add a white to

4.541: this one,
4.542: it'll give me that
4.543: one. Um, I add a white to this one, it'll give me that
4.544: one, and then if I add a white to, um, this
4.545: one, it'll give me this one. And then if I
4.546: add a white to this one, um, it'll give me -- or
4.547:
4.548: so you have two blacks, two whites -- it'll give me that one. Did I already do that one?
4.549:
4.550: I: I think
4.551: [00:25:11] you may have said it twice. [Oops.] But that's okay. /c/
4.552:
4.553: S: I
4.554: was gonna say that sounds familiar. And
4.555: then if I add a white to this one, I'll
4.556: get that one. [OK] And then for blacks,
4.557: adding a black to this one,
4.558: it'll give me that. Adding a black to
4.559: this one will give me that one. Um, adding a
4.560: black to this one, would give me this one.
4.561: Adding a black to this one would give me
4.562: that one. Adding a black to this one would give me that one. Um, adding a black to this one with
4.563:
4.564: [00:25:34] give me that one, and then adding a black
4.565: to this one would give me that one. Um,
4.566: adding a black to this one would give me
4.567: this one, and adding a black to this one, um, that would give me that one.
4.568:
4.569:
4.570: I: Okay. So do you feel satisfied you have
4.571: all the combinations?
4.572:
4.573: S: I think so. /c/

4.574: Growing Possibilities: 0-1 Sample Size Plots

4.575: I: Okay.
4.576: Great. Um, so I'm gonna pick one of the
4.577: samples here in this plot and I'd like

you to draw now, um... {did I give you a pen yet? I did not.} Um, [00:26:14] the sample size plot for that sample. So these are the graphs we made earlier where the mean is on the x-axis /S yawns/ and the sample size increases as you go up the y-axis. So what would the sample size plot look like for -- um, for this -- between N equals 1 and N equals 4 for this sample. So you can draw just up to the dotted line. So for this sample.



S: What the mean is?

I: Yeah, so the mean at each sample size, [00:26:46] going from 1 to 4.

S: Wait wouldn't that just be -- wait I'm confused on what the graph is supposed to be is.

I: Okay. Um, so the graph shows the mean at each sample size [Yeah]. Just like we kind of have the means here. So right now at for the mean is already -- is at 1. Um, and this is a sample where we drew black black black and black, so 1 1 1 1. We drew a one every time for this sample. [00:27:19] So -- um, so you can actually follow backwards if you want. So what was the mean here?

S: One.

I: And then here?

S: One.

I: And then here?

S: One.

I: And then here?

S: One.

I: Yeah. So it was -- the -- the mean was one at every sample -- at every -- everywhere from 1 to 4.

S: Yeah. So how am I supposed to draw like just -- if it's just for the black, like, it would just be a point for N equals 4.

I: Right. It would be a point. But if we draw it at each sample size, we [00:27:55] could draw -- p-- one point for each sample size, right? So at n equals 3, this was all black, which means it was the same as this so it was also at 1.

S: Yeah.

I: So, um -- so --

S: So would I just draw 4 points, then, because this is like N equals 1 and this is N equals 4?

I: Mh. Sure. And you can draw -- you can draw -- or you can draw it as a line, like the way we were before.

S: OK...

I: Um... [00:28:29] so this is just for this sample.

S: Oh, okay.

I: So, um, what do you think, um, might happen as the sample grows from, um, 4 up to 25?

S: For all black?

I: Starting with four black, yeah. What do you think might happen? As this samp-- as we =keep on adding on to sample.=

S: =Umm, the average --= the mean won't change if they're all black.

I: Right. [So...] But we're still randomly drawing and adding on to this, so if we happen to have a sample that started out with all [00:29:07] black, four black. But then we kept on randomly drawing from the box, um, and taking the mean, um, just like we had been doing before in the previous activity. And it's okay if you don't understand something and you're -- [Well...] feel free to ask for clarification. /inaudible/

S: The probability of getting all black will just decrease as you increase the sample size, because it'll be more difficult. [mh] [00:29:34] So...

I: So I'm asking what you think is likely to happen after this. So we've -- we've already seen that the first four are black, but then if we kept on randomly drawing after that, what might happen from there.

S: It can only go left. [OK] Well, I mean, like, if -- like it it can go back to the right but it -- the white will always -- like it'll still shift to the [00:30:00] left. Like, even if there's only one right, it'll go to the left. [OK] It can still go the right, but it'll probably won't -- most likely it probably won't be like at one for another couple of times.

I: Okay. So can you draw what you think might happen to this once we start with all four black.

S: Um...

I: Okay. And can you describe to me what you just drew?

S: Well, each black or each white [00:30:35] will definitely affect it, and it -- it might go like back -- it'll probably go back and forth, um, each time, cuz, like, each time you'll probably get a mixture of black and white. [mh] So.

I: Okay. Great. So, um, we'll just do the same thing with a couple other sort of these samples here. So what would the sample size plot like -- look like from, um, one to four for this block?

S: So...

I:
[00:31:15] For this sample. So it's a sample where we drew white the first time, black the second time, white the third time, and white the fourth time.

S: Like, what are the chances -- I -- I don't know -- I'm confused about like what the graph has supposed to, like -- what the graph is a-- like what you're asking. [OK] Like the graph is supposed to be.

I: Sure.
So we're looking at the graph of, um, what

[00:31:42] the mean is at each sample size. So
this is a sample. So we've drawn four
already where we started out at white.
And we can see that here that would be
at zero.
And then we had -- then we drew a black, so
that's looking like this so far. And then
we drew a white which brings us back
over to here. And then we drew another
white which brings us back to here. [Yeah]
[00:32:15] So we would draw, um, just the means at those --
we would just draw those four means.

S: Okay.
So it started off with zero, and then
it went to one, and then it went to 0.5, and
then it went to -- the -- I guess, like, 0.3, { /inaudible/ I guess...}
And then I went -- or -- it went to point 5 -- two five. [Um...]
Is that what you're asking, or...

I: I think so, so -- but we just
don't have that, um -- we don't have that
stretch to one at all. Because it never
[00:33:13] went -- the mean was never one.

S: Oh, yeah it'd --
it'd be... okay.

I: So you can just cross that
out and you can [Yeah] connect the -- connect to
the 0.5. Okay.
Um, and so we'll just do that same thing
again for, um, this sample, um, where we drew -- [OK] what?

S: So it starts it at one. So for that one? So
then for two it'll stay at one, and then
for three, it'll point six six,
and then for four it will go to point 75.

I: OK. So one thing I forgot to [00:34:08] ask for is could you draw what you think is likely to happen after, um, as we increase the sample size after that point.

S: OK. So it'll be squiggly, and then it'll kind of stay more towards like the 50/50 splitter. [OK] Whatever fraction is pretty close to that.

I: OK, great. And the same thing for this latest block.

S: For the same thing. It'll probably zigzag [00:34:36] around for a little bit, um, but after quite a few trials, it -- or the sample size increases, it should be about a 50/50.

I: Okay. And finally, um, this one here.

S: So the white and then three black? It'll be at zero, and then it'll go to 0.5, and then it will go to 75%, and then one.

I: =So I think it never went to one.=

S: =Or, no. Never one.= Sorry. Point six six. There we go.

I: OK. And again, what you think [00:35:31] would happen after at that point.

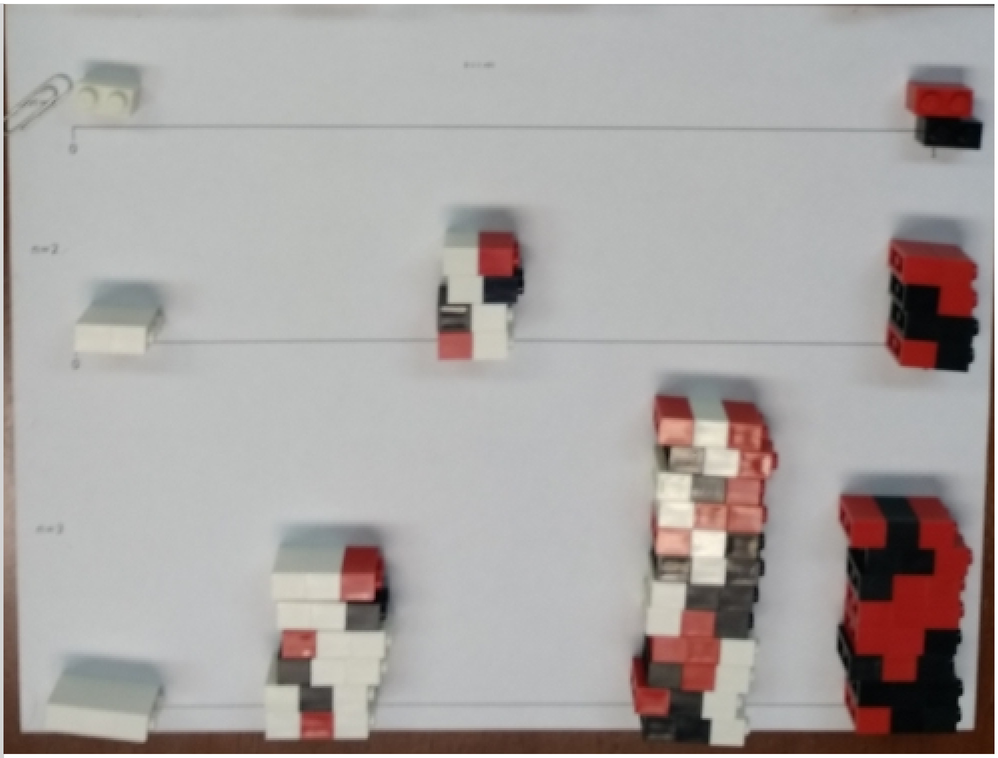
S: Um, it'll probably zigzag around still, and then be somewhere close to 50-50.

I: Okay. And so, um, what do you think will happen -- and so what do you see happening to the distributions as the sample size increases?

S: Um, they get more spread out, but there's just more of each, under -- like for -- there -- cuz there's more -- there's different, like, combinations that you can [00:36:05] get, and there's more of them for each time you increase the sample size. So it'll just keep on increasing, um, the number that you could potential-- like the -- the number of potential combinations that you could get each time you draw [OK] four blocks -- or however many sample size there is.

Growing Possibilities: 0-1-1 Building Blocks

I: So now we're going to -- um, I'll just move this to the side. I never knew I would be professionally dealing with Legos [/c/] but here I am. /c/ Um, so we're gonna add a [00:36:51] red block into the box, um, but the red is also scored is one [OK]. Um, so it's scored just exactly the same as the -- the black block. Um, and just like before we shake the box randomly pick a block and put that block back into the box. Um, so here at, um, n equals 1, uh, what are the possible outcomes?



S: Um, so it would be either the -- the white block, the white and black block -- or -- I mean, sorry -- the black block or the red block.

I: OK. Um, [00:37:47] so and -- um, and what are the means of each of those possible outcomes.

S: Um, one or zero.

I: Okay. Um, so now we want to look at all the possible outcomes, again, when we draw two blocks from the box. Um, so, um -- you can create all of the ones ending in a white block.

S: Ending in a white block. OK. {So then there's that one. And...}

I: Okay. And, um,

again, how did these, um, blocks -- combinations relate to what you saw at n equals one?

S: That there is a higher proportion of the [00:38:46] white blocks, um, and there's -- uh, more combinations that you could end with white.

I: Okay. Um, s-- so let's do all the ones that end in, um, a black block. And, um, let's do all the ones that would end in a red block. Okay. And, uh, what do you notice so far?

S: Um, that there is an equal amount of, um, options for getting like a 50/50, or getting one. [mh]

I: And which means are kind of the most -- would [00:40:08] be the most likely for randomly drawing two blocks from a box?

S: Um, either like a point five or one.

I: Okay. All right. So let's look at -- um, if we're drawing three blocks now. And again let's start with all the ones ending in a white block. So what have you -- what are you looking at to the side -- what -- what to do next?

S: Um, I'm looking at the combinations I made before. [OK] [00:42:05]

I: Okay. Um, so we can do the ones that, um, -- well, I'll just ask. Um, and how do these relate to the possibilities you saw at n equals two?

S: Um, there's more possibilities of getting -- or ending with a white block than there was before [mh], because of the different combinations available.

I: Okay. Great. Um, so you can do all the ones ending in black block.

[00:43:40] Um, and let's do all the ones ending in a red block. Okay. Um, so what do you notice?

S: Um, that you're more likely to get a combination with, um, a -- a white red and black block, or something that -- where the mean equals, um, point seven. [OK] Like, around point, um, six. [OK] Six or so. [And --] Rather than any of the other choices.

I: Okay. [00:45:42] And why do you think that one may be coming out as the top?

S: Um, because the red and black block are both one, um, so if you have a -- like that's basically like a two-to-one, so like you're more likely to be able to get a higher mean rather than a lower mean, because you only have one option for getting like a mean of zero for the combination.

I: Okay. And, uh, what do you think would happen, um -- w-- what would you [00:46:14] expect to happen if we kept increasing the sample size?

S: That the, um -- combination
where it's like the one that's closest,
or m-- it's gonna be more right, on like the
right side of the graph, probably more
towards one than, um, anywhere else, so.

I: OK.
And, um, at say N equals 10, would zero be more
or less likely than here at, um, N equals
fou-- n equals three?

S: Um, less likely just
because the -- uh, like the other probabilities
[00:46:52] are just increasing while that one's just
staying the same, so.

I: Okay. And what about
at one? Would one be more likely or less
likely?

S: Um, I'd say it would get more and
more likely. But, {yeah}, yeah, because it -- the -- the
combinations are only going to increase. Um,
so I would say it's probably gonna
become more [OK]
-- more likely or stay around, like,
the same, like, area. So.

I: Okay. Um, and s-- if I put, um, say a
[00:47:26] divider
somewhere around here. Would you expect
the -- the percentages in that area to increase,
decrease or stay about the same as we =increase the sample size?=

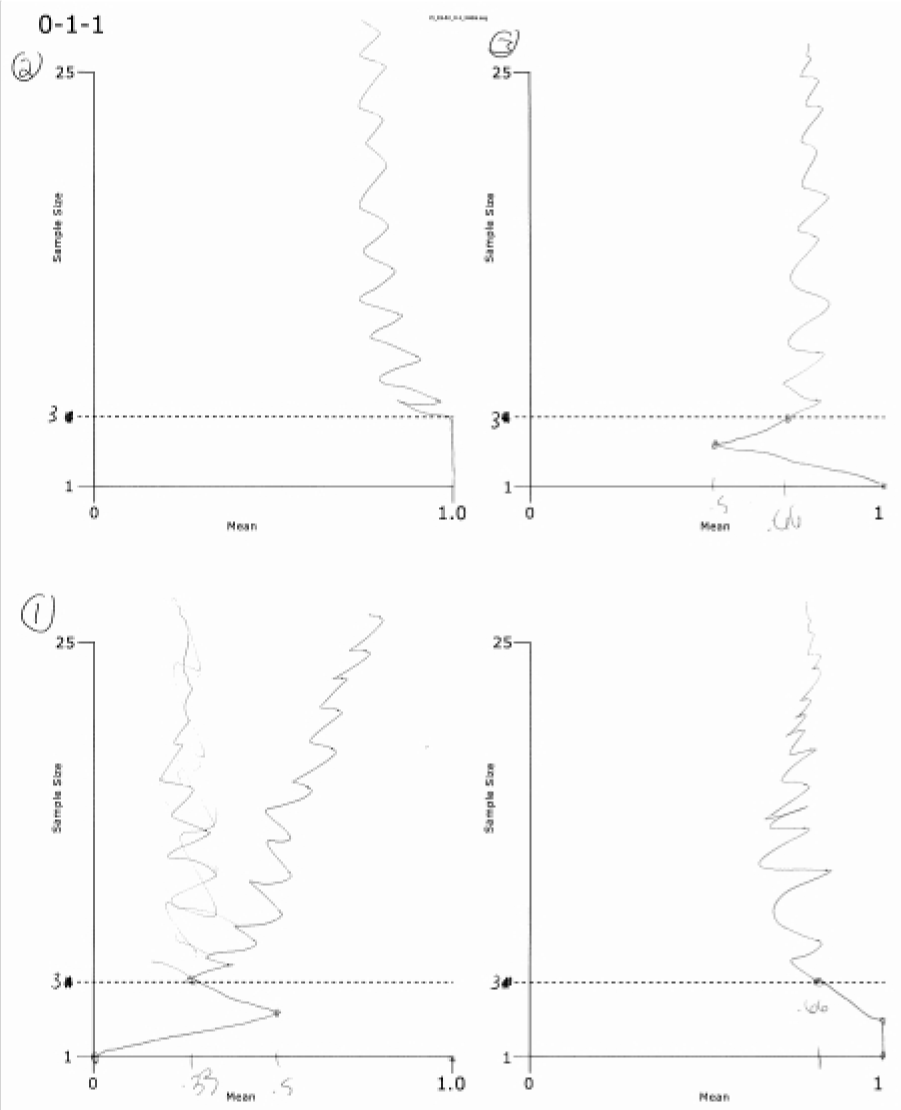
S: =I would say they probab-- I= -- I'd probably say
that they would increase. [OK] So.

I: And why is that?

4.1015:
4.1016: S: Um,
4.1017: especially as you increase the sample
4.1018: size, each fraction will be able to like --
4.1019: they'll be more areas, um, where you can get -- it's like
4.1020: a certain combination. Uh, and so having
4.1021: [00:47:57] that range, like, it's just gonna get more --
4.1022: more of them are just gonna keep on
4.1023: increasing. [OK]
4.1024: So.

4.1025: Growing Possibilities 0-1-1: Sample Size Plots

4.1026: I: Um, OK. So um, I'm going to do the same thing
4.1027: with the -- {let's... let's move this here. Get a little more room.}
4.1028: So we're gonna do the same thing of the sample
4.1029: size plots again. Only now we're only
4.1030: going up to three, since that's what
4.1031: we've gone to here. Um, and, um -- so what would the
4.1032: sample size plot look like, um, for -- for this
4.1033: [00:48:50] sample?



S: So the two reds and then a black?
[Yep.] It would go -- just go u-- straight up.

I: Okay.
And what would happen as you kept on
increasing the sample size.

S: Um, it'll
probably go back and forth in between
somewhere in that area [OK], decrease,
increase, a little bit, so.

I: Okay. And would it keep -- would the -- um, would these back and forth zig zags stay about the same, or would they...

S: [00:49:21] It would get smaller as you go up I would say, probably.

I: OK. Um, and then what would the sample size plot look like for this one?

S: That one. OK. So, it would start at one...

I: Could you just do it up here?

S: Oh yeah. Forgot about that one. /cs/ So it would start at one, and then, um, it would go to 0.5, and then it would go to -- back to around, I'd say 0.66.

I: Okay. And, uh, again draw you think what happen as the sample size increases to 25. [00:50:13] And the same pattern that you were talking about there?

S: Yeah, how the zigzag will just get smaller.

I: Okay. Um, and then for -- um, [/yawns/] {oops}. Oops, I skipped one.

S: That's okay.

I: So I guess this -- we'll two, three, just for my own notes.

S: Yeah, of course.

I: /c/ Um, and this is one. Um -- um -- for... this one.

S: So the white, red, white. So it would start at one, um, and then it would go to a point five, and then it would go to point three three, and then -- for it going up to like 25, you'll probably go back and forth for a while and kind of get s-- like, less [00:51:13] zigzaggy as you go up. [OK] And probably centered somewhere on the left side of, um, the graph. [OK. And...] Or wait. No, that's wrong. [OK] Okay, it would go more towards -- it'll probably end up, um, going more towards like the right, and kind of zigzag more to the right. Just because there's more of like the green -- or not green -- the red or black blocks that would definitely shift it over [OK] to one.

I: So even though it started out on this side, um, [00:51:49] it -- it might still shift over after that.

S: Yeah.

I: Okay. And, um, this one.

S: The black black white? Okay, so then it would start at one, stay at one, and then we go to around point six six. And then it would probably just kind of zigzag and then smaller zigzags as you get up.

I: Okay. So, um, {actually... let's put these in here for right now, just for right now.}

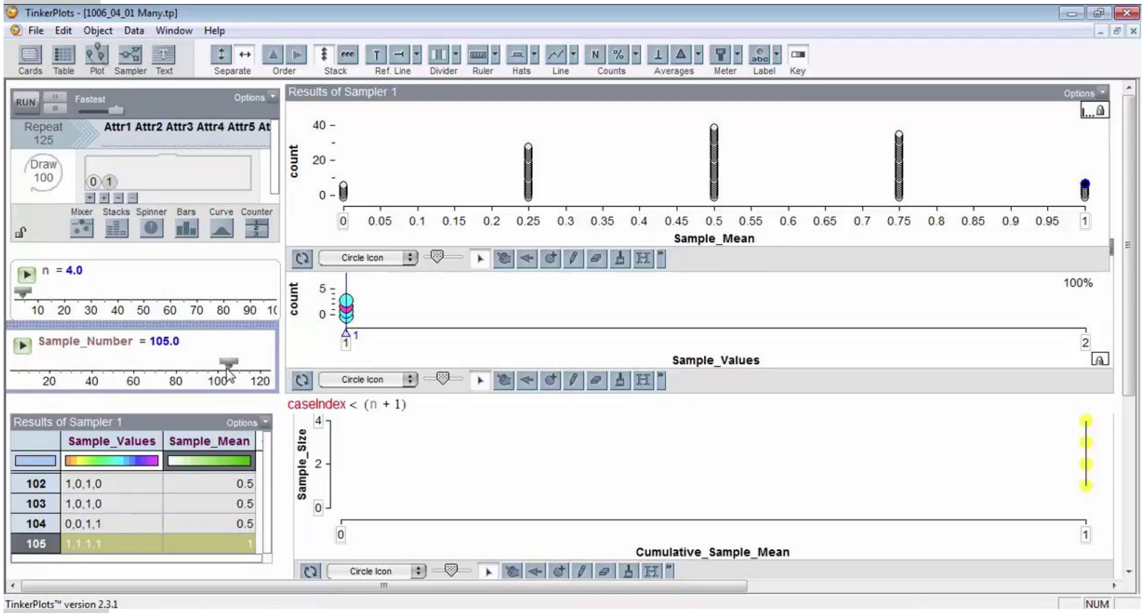
S: Do you need the black ones, too?

4.1126: Growing Many Means: 0-1

4.1127: I: Um, sure.
4.1128: OK.
4.1129: OK.
4.1130: So, um -- so here's a TinkerPlots file to
4.1131: [00:53:09] explore both the -- I'll get this a little closer.
4.1132: To explore both distribution of means
4.1133: and of -- with the sample size plots at the
4.1134: same time. So this is a little bit like
4.1135: what we were looking at last time, where
4.1136: we have the draw here, and the way it's
4.1137: set up is that instead of controlling
4.1138: the number of the sample through here,
4.1139: I actually control it, um, with this slider here.
4.1140: [OK] So this is at N equals 1, and if I
4.1141: [00:53:44] pick a particular sample -- so that's 25,
4.1142: samp-- 125 which I've highlighted up
4.1143: here. Um, if I increase the sample size to 2,
4.1144: I drew a 1 next and so it's still at 1,
4.1145: and now I'm showing that it drew 1
4.1146: again. Um,
4.1147: and then I drew a 0 so it moved over to
4.1148: up there and this is just taking the
4.1149: mean of these values. Um, so does that make sense
4.1150: [Yeah] so far? [mh] And so we've got a few other
4.1151: [00:54:18] things that we can show in this one as
4.1152: well.
4.1153: So, um, this has an ability -- so this was
4.1154: sample number 125 so I can look to see
4.1155: that sample. {125,
4.1156: right at the end.} And it's sort of
4.1157: highlighted in black up there, so I can
4.1158: see where it is. And then this is the --
4.1159: just the -- the sample value, so we have three
4.1160: ones and one zero, just like we see here.
4.1161: [00:54:54] And then we also have the sample size
4.1162: plot, so we state -- we -- the mean was one
4.1163: and then one and then one and then it
4.1164: moved over to 0.75. [OK] Um, so any questions
4.1165: about how this is set up? [Uh-uh] Um, so let's look

back at the sample size plots that you drew. Um, so can you find a sample, um, that matches the beginning of your sample size plot here, one that... yeah. [00:55:33]

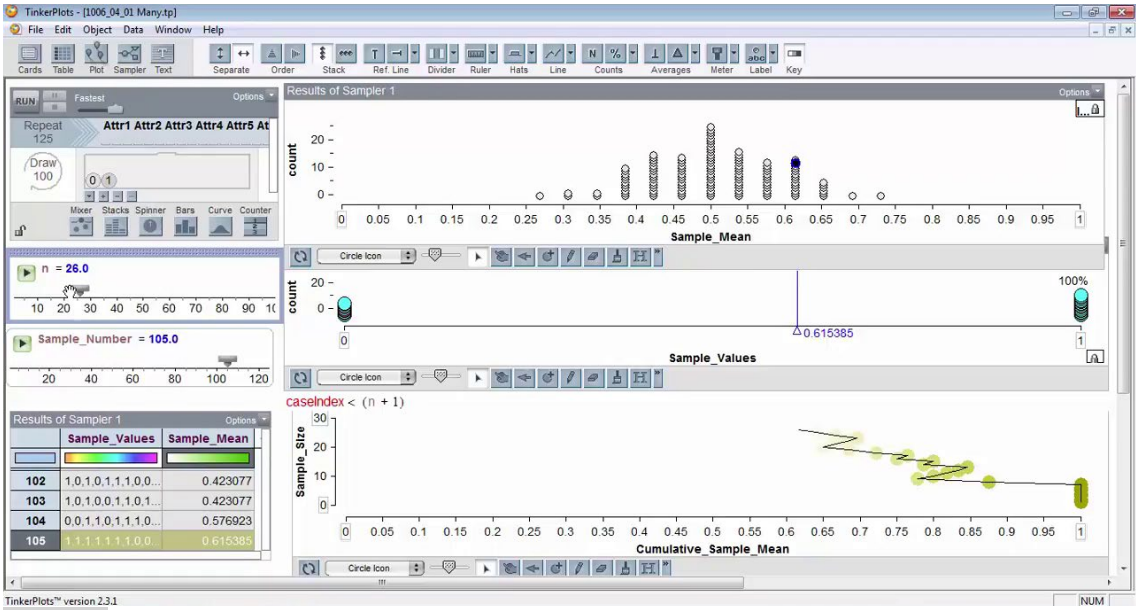
S: Um. Let's see. So it would probably -- at the very top, or... I guess that one, right? Cuz that one's all zeros? [OK] Or -- all ones. Oh. Um. So the 105.



Results of Sampler 1

	Sample_Values	Sample_Mean
102	1,0,1,0	0.5
103	1,0,1,0	0.5
104	0,0,1,1	0.5
105	1,1,1,1	1

I: OK. Um, let's see what happens as it grows to n equals 25. So you can press the -- um -- or I -- I c-- I'll press it [All right] so you don't have to worry about it. I'll press the play button and you can [00:56:18] watch the graphs to see what happens as it grows from 4 to 25. You can watch it move in the upper graph with this black dot, or you can watch the sample size plot below. {Oops.} We're at 26 now. Um, so, um, what do you notice?



Results of Sampler 1

	Sample_Values	Sample_Mean
102	1,0,1,0,1,1,1,0,0...	0.423077
103	1,0,1,0,0,1,1,0,1...	0.423077
104	0,0,1,1,0,1,1,1,0...	0.576923
105	1,1,1,1,1,1,1,0,0...	0.615385

S: That zigzagging. Um, but it looks like it's still going to the left, so.

I: OK, um, and what happened to the plot of means on top?

[00:57:01]

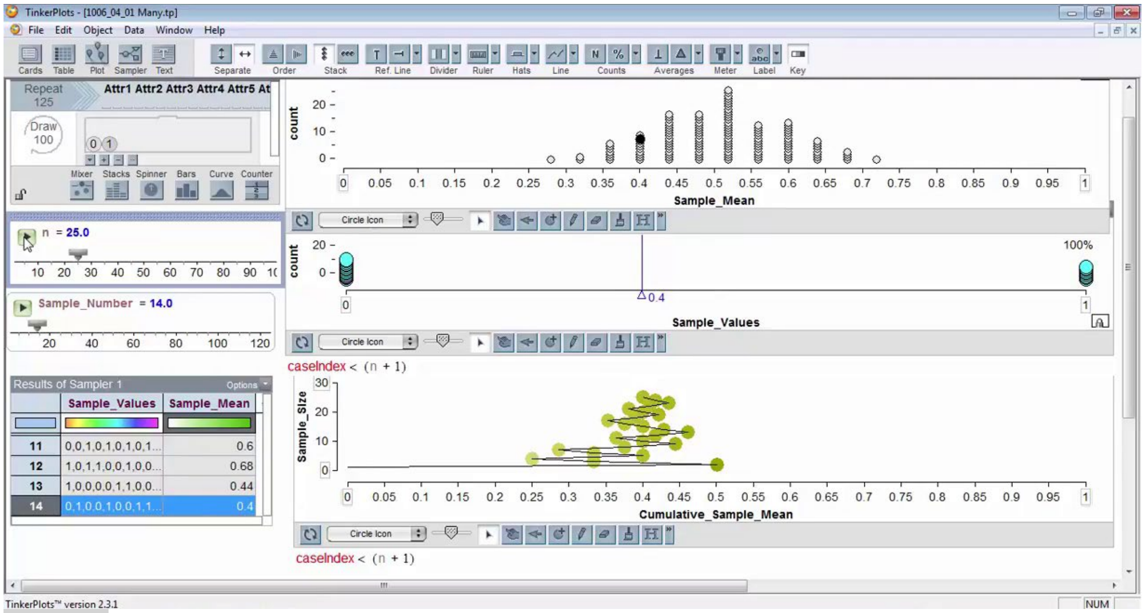
S: Um, it kept shifting either to the -- it kept shifting to the left, and then it'd sometimes go a little bit to the right, but it kept going to the left.

I: Okay. And what about the distribution as a whole.

S: Um, it's a wider distri-- like, there's more -- there's more options.

I: Okay. OK. Um, so let's, um -- um, so let's -- um -- we can -- so go back to N equals four. Actually, we can go down to N equals three and

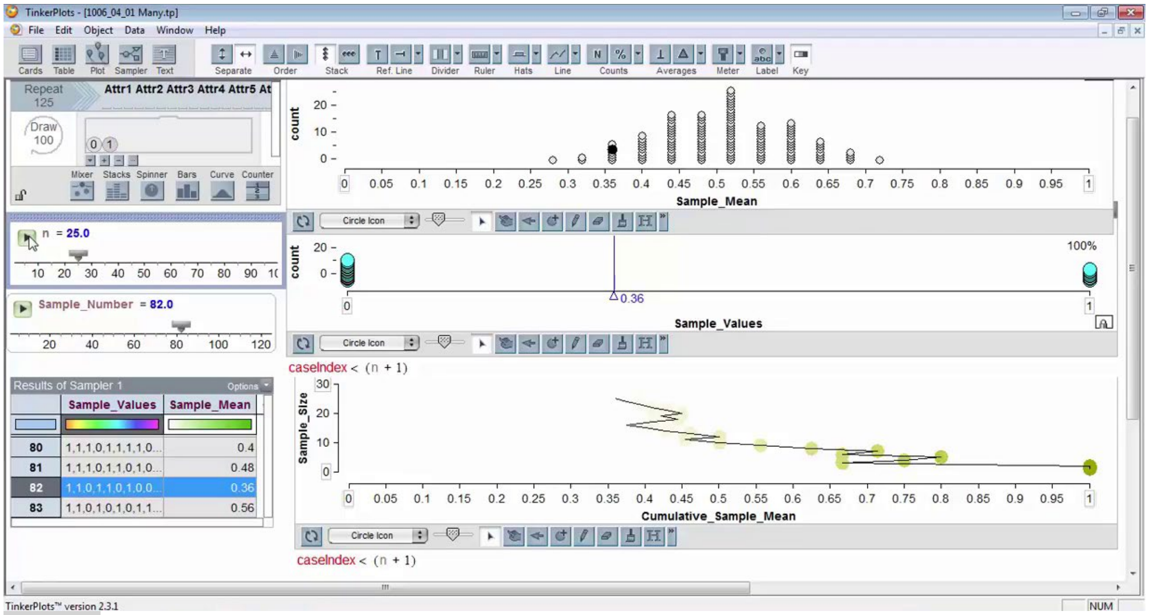
4.1216: [00:57:37] then go back up to four. /c/ It'll just
4.1217: make the slider go the right
4.1218: direction. Um, so can you find a sample
4.1219: that -- /c/
4.1220:
4.1221: S: This is zero, then...
4.1222:
4.1223: I: So this one goes zero, then 0.5,
4.1224: and point --
4.1225:
4.1226: S: =Three three.=
4.1227:
4.1228: I: =Three three= [Yeah] then 0.25.
4.1229:
4.1230: S: So it'd be at zero one
4.1231: zero zero. So, is
4.1232: there, like, a faster way of getting there? /cs/ /S clicks on one of the 0.25 and then scrolls up and down sample/ Zero... So
4.1233: 14...
4.1234:
4.1235:
4.1236: I: Okay.
4.1237: And let's go -- let's, uh, look at that sample.
4.1238:
4.1239: S: So,
4.1240: [00:58:35] do you just -- =is it highlighted right now?=
4.1241:
4.1242: I: =So you move this...= yeah, yeah.
4.1243:
4.1244: S: Oh,
4.1245: yeah. I have to go up down here. OK, so 14.
4.1246:
4.1247: I: I wish I could make it
4.1248: [Yeah /c/] do it so you just highlight, but I
4.1249: couldn't figure out how to do that. Okay.
4.1250: Um, so let's see what happens as this
4.1251: grows to 25. [OK]
4.1252: And so what do you notice so far?
4.1253:


TinkerPlots - [1006_04_01 Many.tp]
File Edit Object Data Window Help
Cards Table Plot Sampler Text Separate Order Stack Ref. Line Divider Ruler Hats Line Counts Averages Meter Label Key
Repeat 125
Draw 100
Attr1 Attr2 Attr3 Attr4 Attr5 At
Mixer Stacks Spinner Bars Curve Counter
count
Sample_Mean
n = 25.0
Sample_Values
100%
0.4
Sample_Number = 14.0
Circle Icon
caseIndex < (n + 1)
Sample_Size
Cumulative_Sample_Mean
TinkerPlots™ version 2.3.1
NUM

Results of Sampler 1

	Sample_Values	Sample_Mean
11	0,0,1,0,1,0,1,0,1...	0.6
12	1,0,1,1,0,0,1,0,0...	0.68
13	1,0,0,0,0,1,1,0,0...	0.44
14	0,1,0,0,1,0,0,1,1...	0.4

4.1254: S: Um, it kind
4.1255: of went to the right, and then it's zig
4.1256: zagging. Seems like it's not going a
4.1257: certain direction per se, it's kind of
4.1258: [00:59:17] just like staying somewhere in the
4.1259: middle of zero and one, and then [OK] the top
4.1260: graph, it kind of was just going back and
4.1261: forth, like, um, kind of like how this was, it
4.1262: was just kind of going back and forth.
4.1263: Wasn't varying a whole ton, so.
4.1264:
4.1265: I: Okay. Um, so we'll go
4.1266: back to 4.
4.1267: And, um, let's find a sample that matches
4.1268: this one.
4.1269:
4.1270: S: So one one zero one.
4.1271: So eighty two. {One one zero one.}
4.1272: [01:00:15]
4.1273:
4.1274: I: Okay. Let's see what happens as we -- as it
4.1275: grows to n equals 25.
4.1276: So what do you notice so far?
4.1277:


TinkerPlots - [1006_04_01 Many.tp]
File Edit Object Data Window Help
Cards Table Plot Sampler Text Separate Order Stack Ref. Line Divider Ruler Hats Line Counts Averages Meter Label Key
Repeat 125
Draw 100
Attr1 Attr2 Attr3 Attr4 Attr5 At
Mixer Stacks Spinner Bars Curve Counter
count
Sample_Mean
Circle Icon
100%
0.36
Sample_Values
caseIndex < (n + 1)
Sample_Size
Cumulative_Sample_Mean
caseIndex < (n + 1)
TinkerPlots™ version 2.3.1
NUM

n = 25.0

Sample_Number = 82.0

Results of Sampler 1

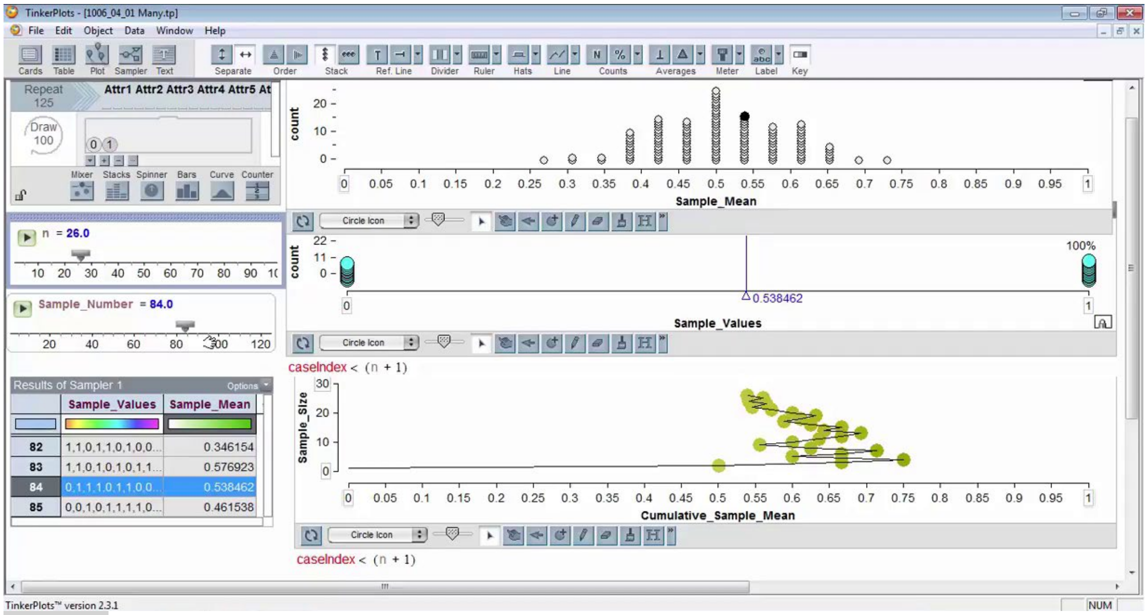
	Sample_Values	Sample_Mean
80	1,1,1,0,1,1,1,1,0...	0.4
81	1,1,1,0,1,1,0,1,0...	0.48
82	1,1,0,1,1,0,1,0,0...	0.36
83	1,1,0,1,0,1,0,1,1...	0.56

S: It kind of just like went -- shot to the left, and [Hm] I mean it was zig zagging but it definitely got -- you can tell that there was a lot of o-- zeros in a row [Hm], so it made it shift quite a bit.

I: Mh. All right, and finally, um, a sample that matches this one. Oh -- and go -- I'll take it -- take you back to n = 4 again.

S: So it'd be zero one one one...
[01:01:18] So, eighty-four

I: So it's highlighted up there now.
Let's see what happens.
Oops. Well, we went to 26. /c/



n = 26.0

Sample_Number = 84.0

Results of Sampler 1

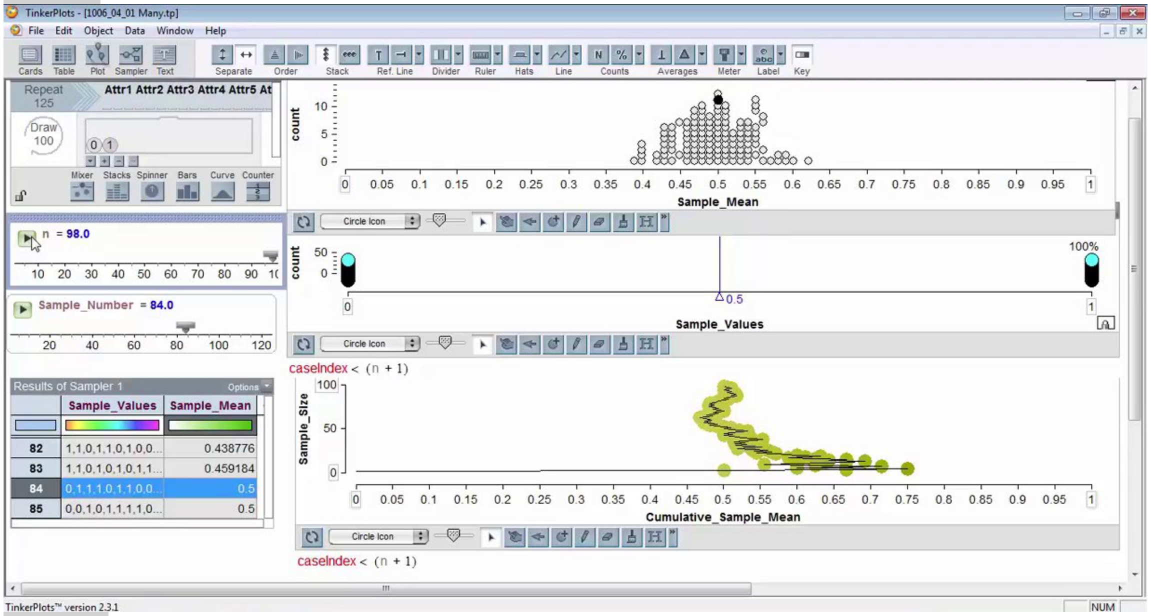
	Sample_Values	Sample_Mean
82	1,1,0,1,1,0,1,0,0...	0.346154
83	1,1,0,1,0,1,0,1,1...	0.576923
84	0,1,1,1,0,1,1,0,0...	0.538462
85	0,0,1,0,1,1,1,1,0...	0.461538

4.1297: S: That's OK. And then, um, so the graph on
4.1298: this one it kind of went all the way to
4.1299: the right, and then kind of stayed on the
4.1300: right side of, um, it being like a 50/50, and
4.1301: kind of s-- went back and forth and stayed
4.1302: pretty -- basically in between, like, I'd say .5 and
4.1303: 0.75, it didn't really go outside of that
4.1304: [01:02:08] range. And then for the, um, the mean with all
4.1305: like the little dots on the top, it kind
4.1306: of just went back and forth between, like, um,
4.1307: those same numbers.
4.1308:
4.1309: I: Okay. What do you
4.1310: think would happen if we grew this
4.1311: sample all the way to N equals 100?
4.1312:
4.1313: S: Um, I
4.1314: think it'll get centered around probably
4.1315: in the middle, I would say. Um, more towards
4.1316: like 0.5 [OK] if you do a lot, because there's
4.1317: a 50/50 chance of you getting it, so [OK] the
4.1318: [01:02:39] more you do it, the more close it'll be
4.1319: probably to that.
4.1320:
4.1321: I: Okay. Let's see what happens

So what do you notice happening so far?

S: Um, it's moving less and less. [OK] At least on the top graph, like it -- because there's more -- different, like, fraction numbers that you can get, um, [Hm] so it's definitely -- there's l-- it's not moving quite as much back and forth.

I: Okay. Um, so tell me a little bit more about [01:03:30] that. You're saying it's not moving as much back and forth because there's more fraction numbers than y-- that you can get?


TinkerPlots - [1006_04_01 Many.tp]
File Edit Object Data Window Help
Cards Table Plot Sampler Text Separate Order Stack Ref. Line Divider Ruler Hats Line Counts Averages Meter Label Key
Repeat 125
Draw 100
Attr1 Attr2 Attr3 Attr4 Attr5 At
Mixer Stacks Spinner Bars Curve Counter
count
Sample_Mean
Sample_Values
100%
0.5
Cumulative_Sample_Mean
Sample_Size
Circle Icon
TinkerPlots™ version 2.3.1
NUM

n = 98.0

Sample_Number = 84.0

caseIndex < (n + 1)

Results of Sampler 1 Options

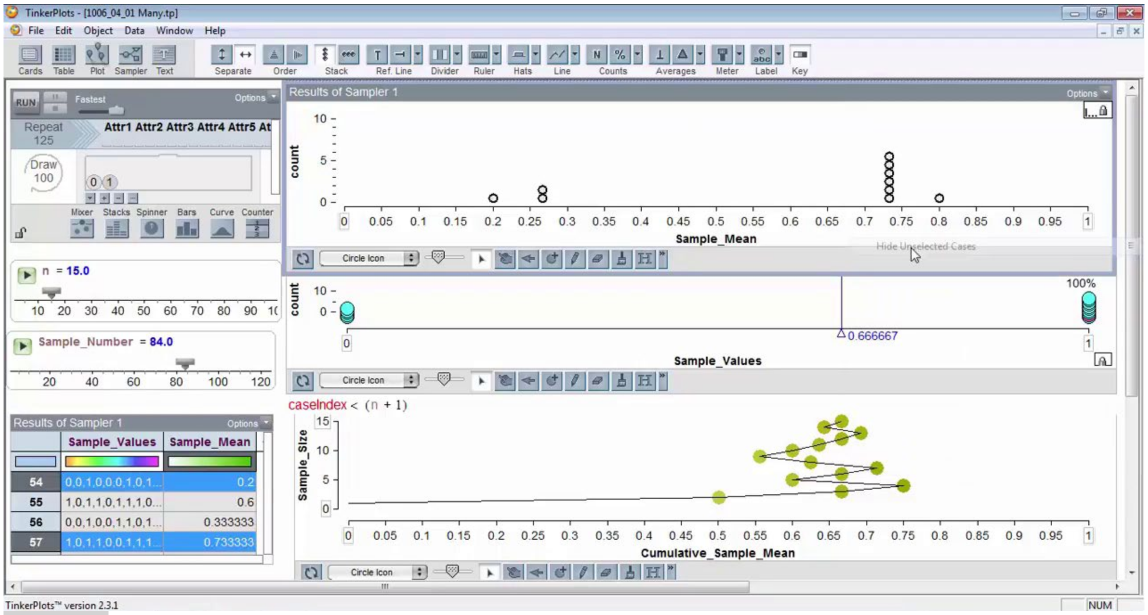
	Sample_Values	Sample_Mean
82	1,1,0,1,1,0,1,0,0...	0.438776
83	1,1,0,1,0,1,0,1,1...	0.459184
84	0,1,1,1,0,1,1,0,0...	0.5
85	0,0,1,0,1,1,1,1,0...	0.5

S: Yeah, just cuz there's more options -- there's more combinations for each thing, so, like, say, like you have only four -- like you have -- you can choose four blocks, like that's -- like you have one out of four chances [mh], like, or like -- it's like 25% increments, or 0.25 increments? While, like, [01:03:52] if you only have, like -- um, like three, it'll be like 0.34. [OK] And so it just kind of keeps on getting smaller -- like the spots -- like, in between those numbers. [OK]

I: Um, and, um -- so that's -- you're using that to explain that graph or this graph?

S: The top graph. [OK] Um, just because it's not moving as much, so. [OK] It's kind of like -- if it do-- like if one block of the -- of a different color gets drawn, like it's not gonna [01:04:25] it's gonna mo-- it's gonna less in one direction, um, than it would, say, like, uh -- in if you only had, like, two blocks. Because there's -- like a higher, like, percentage of blocks in it, so, it's not gonna affect it quite as much. [OK]

I: Um, so, um, we'll do one other thing here. So, um, we have some samples that are kind of towards the extremes here. So here at -- let's say I'll grab a couple of these and a couple of these. Um, what do you think will -- and I'll [01:05:12] -- I can show just those samples. So, uh, what do you think will happen to these samples on the extremes as the sample size grows?


TinkerPlots - [1006_04_01 Many.tp]
File Edit Object Data Window Help
Cards Table Plot Sampler Text Separate Order Stack Ref. Line Divider Ruler Hats Line Counts Averages Meter Label Key
RUN Fastest Options
Repeat 125
Draw 100
Attr1 Attr2 Attr3 Attr4 Attr5 At
Mixer Stacks Spinner Bars Curve Counter
Results of Sampler 1
count
Sample_Mean
Hide Unselected Cases
100%
0.666667
Sample_Values
caseIndex < (n + 1)
Sample_Size
Cumulative_Sample_Mean
Circle Icon
TinkerPlots™ version 2.3.1
NUM

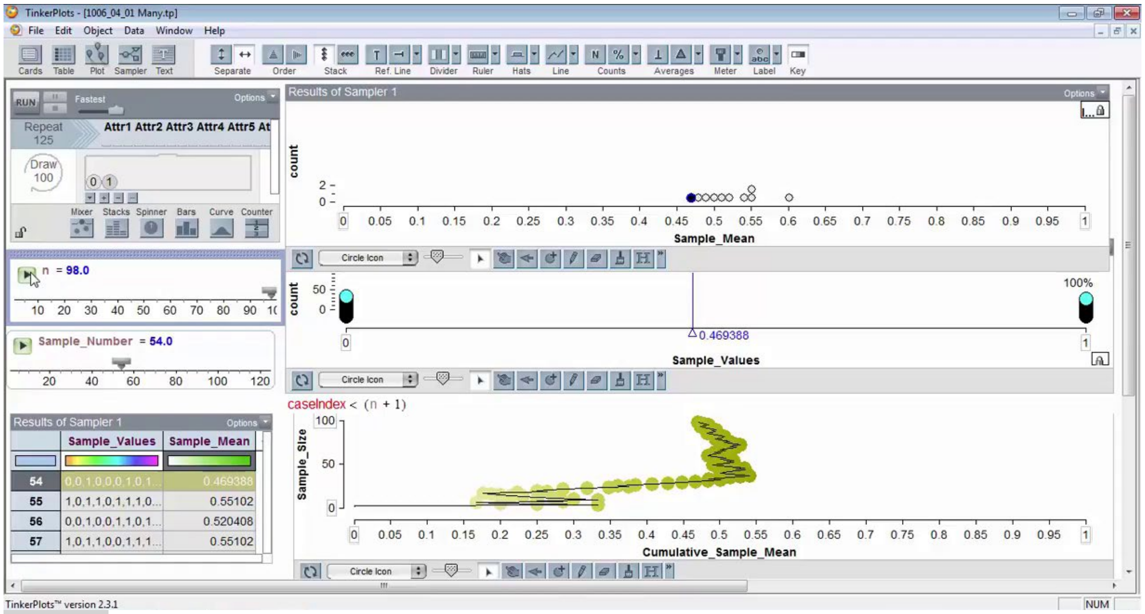
n = 15.0

Sample_Number = 84.0

Results of Sampler 1

	Sample_Values	Sample_Mean
54	0,0,1,0,0,0,1,0,1...	0.2
55	1,0,1,1,0,1,1,1,0...	0.6
56	0,0,1,0,0,1,1,0,1...	0.333333
57	1,0,1,1,0,0,1,1,1...	0.733333

4.1377: S: Um, I think it'll be more
4.1378: difficult to get towards 0.5, but it
4.1379: might take like a little bit longer for
4.1380: it to get more like -- like more centered, I
4.1381: would say. [OK] But, it'll still go
4.1382: towards like a 50/50, so.
4.1383:
4.1384: I: Okay. Um, which, uh --
4.1385: so let's follow one of these as it grows.
4.1386: [01:05:49] Uh, which, uh, which one do you want to look at?
4.1387:
4.1388:
4.1389: S: Let's do point two.
4.1390:
4.1391: I: OK. So that's 54.
4.1392: So this is the sample size plot so far, and let's see. Oops it's going the wrong way. /cs/
4.1393: There we go.
4.1394: Okay. Um, so describe what happened.
4.1395:


TinkerPlots - [1006_04_01 Many.tp]
File Edit Object Data Window Help
Cards Table Plot Sampler Text Separate Order Stack Ref. Line Divider Ruler Hats Line Counts Averages Meter Label Key
RUN Fastest Options
Attr1 Attr2 Attr3 Attr4 Attr5 At
0 1
Mixer Stacks Spinner Bars Curve Counter
Results of Sampler 1
count
Sample_Mean
Circle Icon
100%
Sample_Values
Sample_Size
Cumulative_Sample_Mean
0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5 0.55 0.6 0.65 0.7 0.75 0.8 0.85 0.9 0.95 1
NUM

Repeat 125

Draw 100

n = 98.0

Sample_Number = 54.0

0.469388

caseIndex < (n + 1)

Results of Sampler 1

	Sample_Values	Sample_Mean
54	0,0,1,0,0,0,1,0,1...	0.469388
55	1,0,1,1,0,1,1,1,0...	0.55102
56	0,0,1,0,0,1,1,0,1...	0.520408
57	1,0,1,1,0,0,1,1,1...	0.55102

TinkerPlots™ version 2.3.1

4.1396: S: Um, well we -- we
4.1397: got a lot of blacks in a row. /cs/ And so it
4.1398: went very quickly, um, like to 0.5, it still
4.1399: took probably longer -- it would have had
4.1400: more like black blocks in the very
4.1401: beginning, but, um, having so many in a row
4.1402: [01:07:21] definitely sped it's, uh, way up to getting [/c/]
4.1403: aro-- like centering around, um, 0.5. And then
4.1404: it kind of just stayed in that area and
4.1405: zig zagged around.
4.1406:
4.1407: I: And so, um, what sort of
4.1408: happened to -- kind of, because we had a
4.1409: bunch of other kind of extreme values
4.1410: before. What happened to those as well?
4.1411:
4.1412:
4.1413: S: Um, they st-- definitely -- s-- they it -- they took
4.1414: different amount of times probably to
4.1415: get to 0.5, but, um, they probably -- they're all
4.1416: [01:07:52] going to be centering around 0.5 at some
4.1417: point, um, um, they're like drawing, I guess. [OK] So.
4.1418:

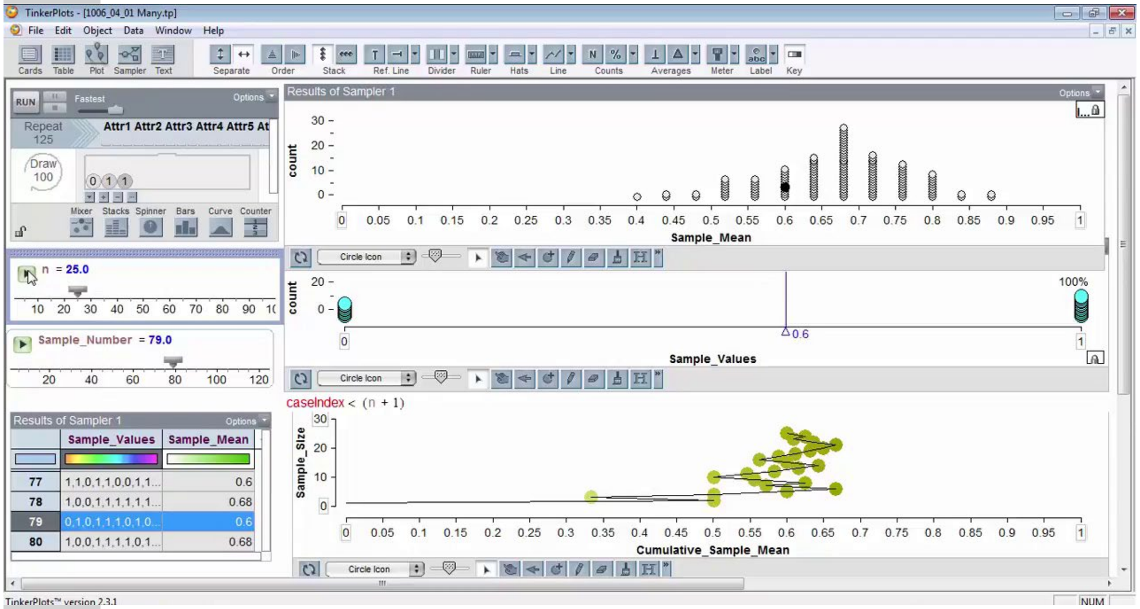
Growing Many Means: 0-1-1

I: Um, okay. So, um, now we're gonna add another one into the sampler. So it'll be more like this situation. [mh] Draw a new sample with those guys. Um, go back down, /c/ to a lower sample size. Um, so just do a couple of these. Um, so let's look at -- can you find a sample that matches -- oh it should be three, huh. Um, can you find a [01:08:58] sample that matches this sample size plot here?

S: Yeah. So /inaudible/ one -- or to zero one one -- or zero one zero.

OK, so then 79, { /inaudible/}

I: Okay. And let's see what happens. Was that what you expected?


TinkerPlots - [1006_04_01 Many.tp]
File Edit Object Data Window Help
Cards Table Plot Sampler Text Separate Order Stack Ref. Line Divider Ruler Hats Line Counts Averages Meter Label Key
RUN
Fastest
Options
Attr1 Attr2 Attr3 Attr4 Attr5 At
0 1 1
Mixer Stacks Spinner Bars Curve Counter
Results of Sampler 1
count
30
20
10
0
0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5 0.55 0.6 0.65 0.7 0.75 0.8 0.85 0.9 0.95 1
Sample_Mean
Circle Icon
100%
0.6
Sample_Values
caseIndex < (n + 1)
Sample_Size
Cumulative_Sample_Mean
TinkerPlots™ version 2.3.1
NUM

Repeat 125

Draw 100

n = 25.0

Sample_Number = 79.0

Results of Sampler 1

	Sample_Values	Sample_Mean
77	1,1,0,1,1,0,0,1,1...	0.6
78	1,0,0,1,1,1,1,1,1...	0.68
79	0,1,0,1,1,1,0,1,0...	0.6
80	1,0,0,1,1,1,1,0,1...	0.68

S: Um, /let's see.?/ [OK] It zig zagged and then kind of centered more on the right side of 0.5.

I: Okay. And why -- why do you think that happened?

S: [01:10:02] Um, why, like, it got -- like went more towards the right? Because there's one more uh-- one in the mixture, uh, that you could potentially get, so the probability of you grabbing like a red block or like getting two one is a lot -- is like -- it's a higher probability than say like the previous one, where it was like a 50/50 shot, so. [OK] You have more -- you have a bit more chance getting a 1 than getting a 0. [mh]

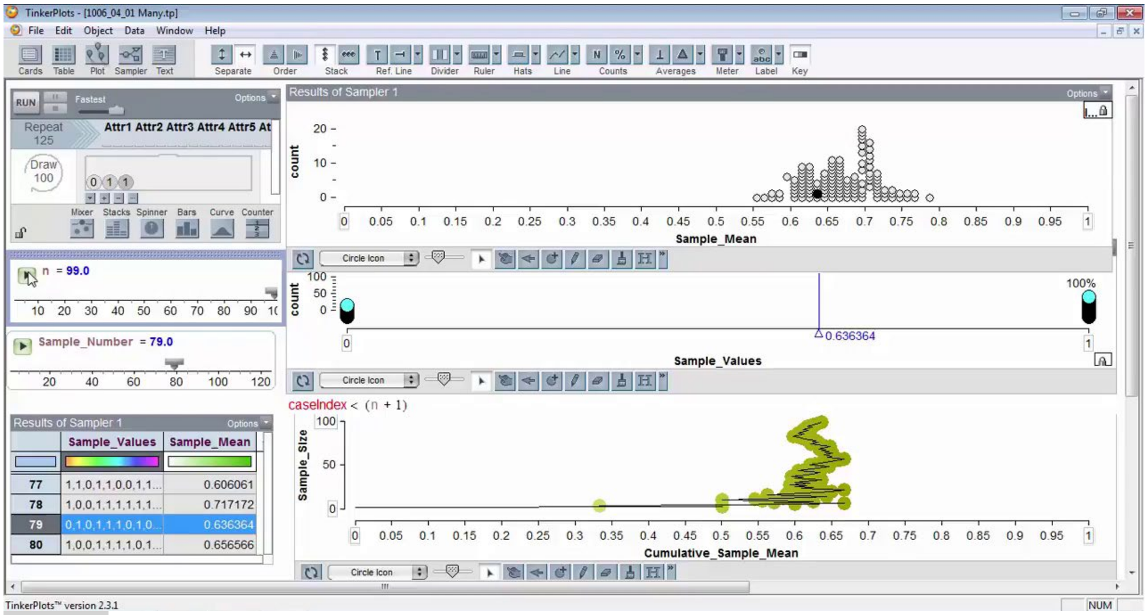
I: Okay. And [01:10:34] what would -- what do you think will happen if we keep on going up all the way to 100?

S: Um, it'll center somewhere in between point five and point one. [mh]

I: And what do you think will happen to this upper graph?

S: Um, kind of like the same thing before. It'll state more towards like one, and then -- then it'll stop jumping, so-- like at large -- it'll be a smaller fraction each time that each time that it jumps to a different area, so. [01:10:59]

I: Okay. So let's -- let's keep going. Okay. So what do you notice here?

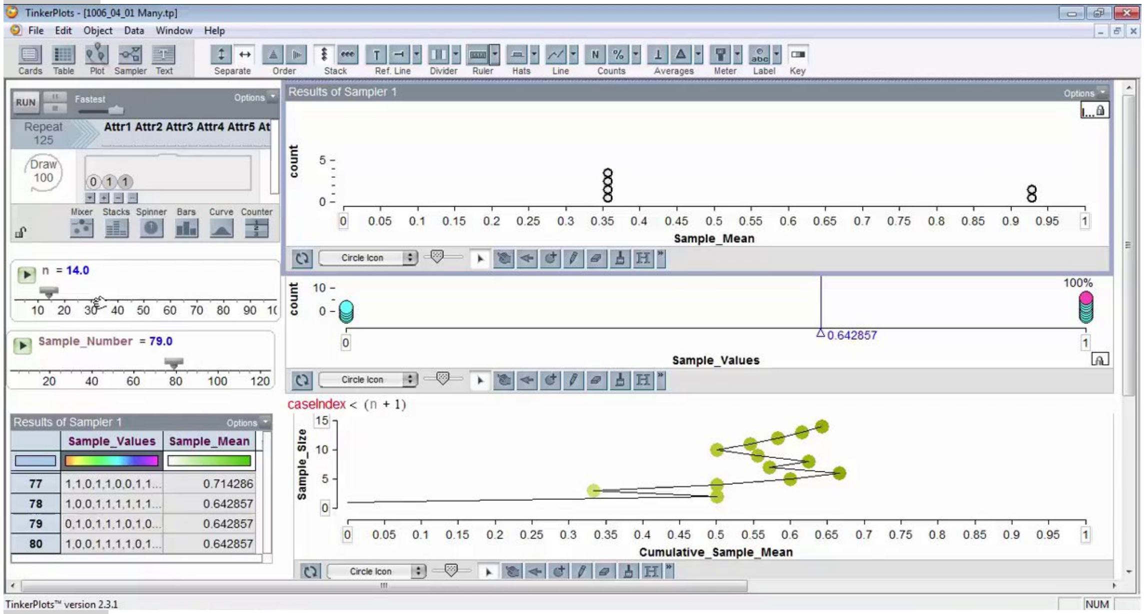

TinkerPlots - [1006_04_01 Many.tp]
File Edit Object Data Window Help
Cards Table Plot Sampler Text Separate Order Stack Ref. Line Divider Ruler Hats Line Counts Averages Meter Label Key
RUN Fastest Options
Repeat 125
Draw 100
Attr1 Attr2 Attr3 Attr4 Attr5 At
Mixer Stacks Spinner Bars Curve Counter
Results of Sampler 1
count
Sample_Mean
Sample_Values
0.636364
100%
n = 99.0
Sample_Number = 79.0
caseIndex < (n + 1)
Sample_Size
Cumulative_Sample_Mean
Circle Icon
TinkerPlots™ version 2.3.1
NUM

Results of Sampler 1

	Sample_Values	Sample_Mean
77	1,1,0,1,1,0,0,1,1...	0.606061
78	1,0,0,1,1,1,1,1,1...	0.717172
79	0,1,0,1,1,1,0,1,0...	0.636364
80	1,0,0,1,1,1,1,0,1...	0.656566

4.1481: S: Um, that
4.1482: it's centered around, like, somewhere in
4.1483: between point six and point six five. [mh] And
4.1484: it kind of just like stayed there. It
4.1485: zig zagged a little bit, but mainly stayed in that
4.1486: area, some -- of -- for the majority of the
4.1487: time after probably like, uh, I don't know,
4.1488: like, 30ish trials or so, [OK] looks like.
4.1489:
4.1490: I: And
4.1491: how did the distribution overall change,
4.1492: [01:12:12] the whole thing?
4.1493:
4.1494: S: Um, there was just more s-- like
4.1495: opportunities like in between each
4.1496: fraction, like, [mh] to get that. Um, but it kind of
4.1497: just like stayed the same and then it jumped
4.1498: around less and less.
4.1499:
4.1500: I: Okay. And I meant -- I was
4.1501: talking about the shape of this -- of all
4.1502: of the means [Oh] together.
4.1503:
4.1504: S: Um, it kind of -- it
4.1505: varied on each one, but it was mainly

4.1506: like, um, like a bell curve, like it kind of
4.1507: like tapered off at -- at the ends, and [OK]
4.1508: [01:12:41] centered around the same area, so. [OK]
4.1509:
4.1510:
4.1511: I: So we'll just did that same thing, looking
4.1512: at extremes.
4.1513: Um, so I'll just select here, and here. Um,
4.1514: so what do you think will happen, um, to
4.1515: these as the sample size grows?
4.1516:



TinkerPlots - [1006_04_01 Many.tp]
File Edit Object Data Window Help
Cards Table Plot Sampler Text Separate Order Stack Ref. Line Divider Ruler Hats Line Counts Averages Meter Label Key
RUN Fastest Options
Repeat 125
Draw 100
Attr1 Attr2 Attr3 Attr4 Attr5 At
Mixer Stacks Spinner Bars Curve Counter
Results of Sampler 1
count
Sample_Mean
0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5 0.55 0.6 0.65 0.7 0.75 0.8 0.85 0.9 0.95 1
Circle Icon
100%
0.642857
Sample_Values
caseIndex < (n + 1)
Sample_Size
Cumulative_Sample_Mean
TinkerPlots™ version 2.3.1
NUM

n = 14.0

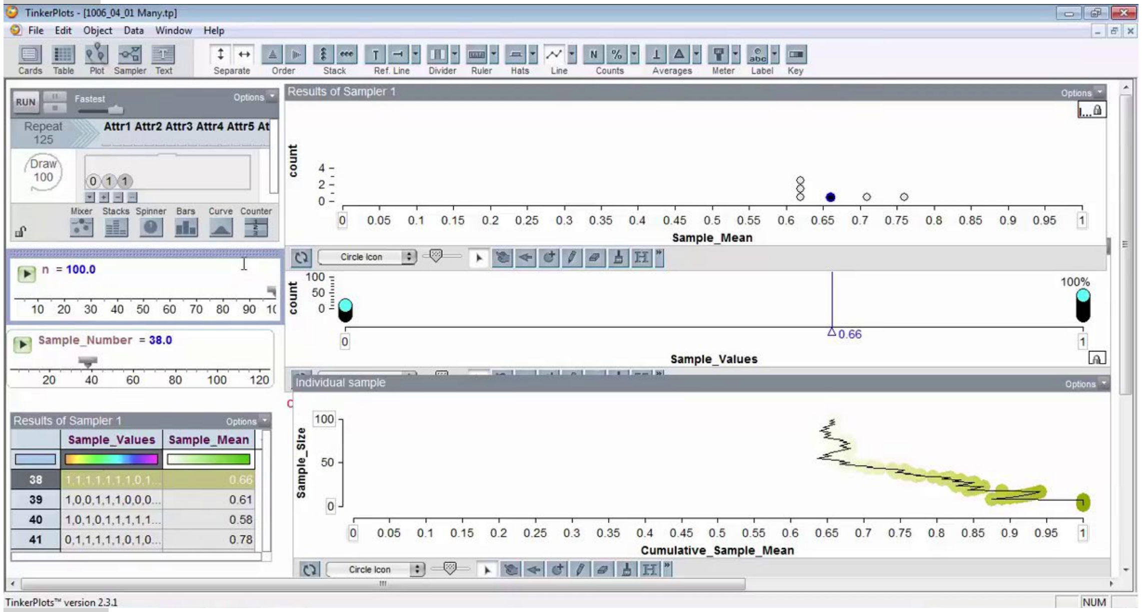
Sample_Number = 79.0

Results of Sampler 1

	Sample_Values	Sample_Mean
77	1,1,0,1,1,0,0,1,1...	0.714286
78	1,0,0,1,1,1,1,1,1...	0.642857
79	0,1,0,1,1,1,0,1,0...	0.642857
80	1,0,0,1,1,1,1,0,1...	0.642857

4.1517: S: Um, they'll
4.1518: definitely still center at some point
4.1519: around like point six and point six
4.1520: five.
4.1521:
4.1522: I: Okay. Well, I see we're running out of
4.1523: time, so I will speed this up. Which one
4.1524: [01:13:25] do you want to look at?
4.1525:
4.1526: S: Um, eith--
4.1527: any one, it doesn't matter to me.
4.1528:
4.1529: I: /c/ Okay. /I picks at around 0.925/ And
4.1530: point nine -- oh, thirty eight. {Oops.} I'm gonna

4.1531: speed this along a little bit faster than -- just... {Oops} /I drags slider up manually/
4.1532: So what did you notice happening?
4.1533:



TinkerPlots - [1006_04_01 Many.tp]

Repeat 125

Draw 100

n = 100.0

Sample_Number = 38.0

Results of Sampler 1

	Sample_Values	Sample_Mean
38	1,1,1,1,1,1,1,0,1...	0.66
39	1,0,0,1,1,1,0,0,0...	0.61
40	1,0,1,0,1,1,1,1,1...	0.58
41	0,1,1,1,1,1,0,1,0...	0.78

TinkerPlots™ version 2.3.1

4.1534: S: Um, so
4.1535: for the top graph,
4.1536: all of them started like slowly coming
4.1537: together. They would jump around a little
4.1538: bit, but they all slowly, um, started going
4.1539: towards like the point six, point six
4.1540: [01:14:21] five area, or point seven, I guess. [mh] And
4.1541: then for the bottom part, it zigzags but
4.1542: it kept c-- shifting more and more towards
4.1543: the left, [OK] um, centered around point six
4.1544: five it looks like.
4.1545:
4.1546: I: Okay. And, uh, one last time, so
4.1547: what, uh -- why do you think that is? Why did
4.1548: it do that?
4.1549:
4.1550: S: Um, probably because it had
4.1551: more, uh, white in the mix, um, and there's a
4.1552: higher, like -- just because it started off
4.1553: with, um, all blacks that doesn't ma-- necessarily
4.1554: [01:14:56] mean that, uh -- or, like, quite a few blacks, that --

4.1555: it doesn't mean that it can't go to the
4.1556: left, um, and it'll know definitely, like -- each
4.1557: white block will, like, affect it too, so. [mh]
4.1558: It won't-- it'll be close to one, but not
4.1559: exactly one.
4.1560:
4.1561: I: OK. OK. Um, that's all for
4.1562: today. [OK]

Appendix D5: Post-Interview

There were technical issues with the screen recording at the beginning of the Post-Interview, so the first three segments below are transcribed from the video recording. Times are relative to the beginning of whichever recording was transcribed rather than to the beginning of the interview.

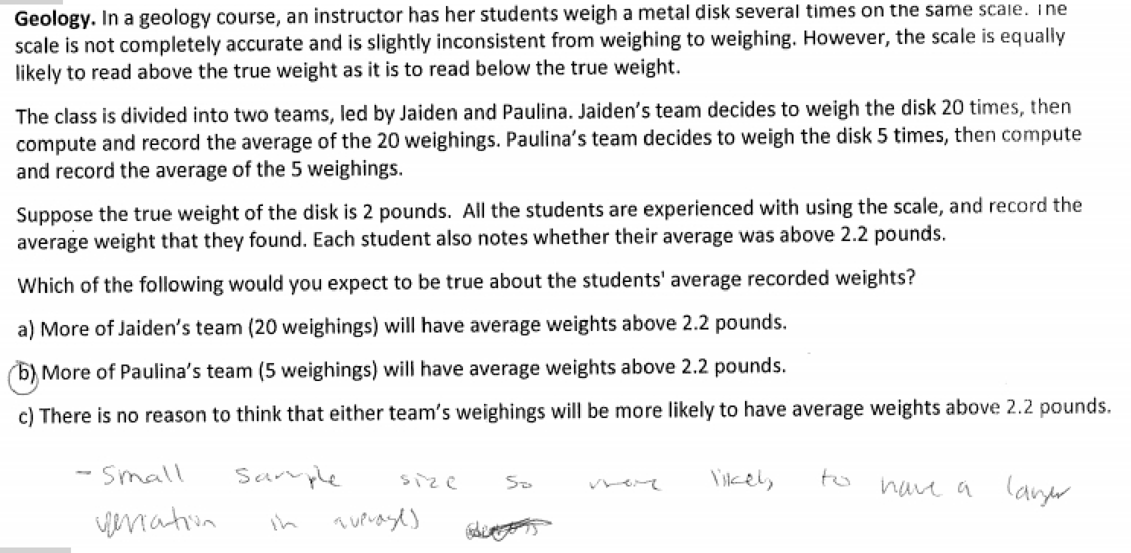
5.1: Post-questions introduction (from videorecording)

5.2: I:
5.3: [00:00:37] Okay. So today, again,
5.4: we're going to -- actually, I can just put this... { /inaudible/ I was recording there.} I'll open it up. { /inaudible/ }
5.5:
5.6: S: That's OK.
5.7:
5.8:
5.9: I: /inaudible/
5.10: /inaudible/
5.11: Because I never tried suspending it while the screen was recording before.
5.12:
5.13: S: Oh,
5.14: yeah.
5.15:
5.16: I: So it got a little confused about what it was doing.
5.17:
5.18: S: Make sense. [/c/] Life is hard. /cs/
5.19:
5.20: I: We ask computers to do so many things /more technical issues: if you didn't like it's
5.21: hard to do so many things as the key
5.22: like they did like recording a screen

5.23: takes so much like link to hold on power
5.24: let me restart anything at starting I've
5.25: [00:02:14] got another please you'll restart
5.26: computer exchange definitely rolling/
5.27:
5.28:
5.29: I: Okay. [OK] So let's start. Um, so once again
5.30: we're going to be solving some problems
5.31: today. So just like on the first day, I'd
5.32: like you to read the problem aloud and
5.33: to think aloud when you answer each
5.34: question. To think and talk aloud as much
5.35: as possible. I may remind you to keep
5.36: talking to keep the conversation moving. Um, any
5.37: [00:02:49] questions before we begin?
5.38:
5.39:
5.40: S: Nope. /c/
5.41:
5.42: I: Okay. Um, we will start with the
5.43: first question.

5.44: Geology (from videorecording)

5.45:



Geology. In a geology course, an instructor has her students weigh a metal disk several times on the same scale. The scale is not completely accurate and is slightly inconsistent from weighing to weighing. However, the scale is equally likely to read above the true weight as it is to read below the true weight.

The class is divided into two teams, led by Jaiden and Paulina. Jaiden's team decides to weigh the disk 20 times, then compute and record the average of the 20 weighings. Paulina's team decides to weigh the disk 5 times, then compute and record the average of the 5 weighings.

Suppose the true weight of the disk is 2 pounds. All the students are experienced with using the scale, and record the average weight that they found. Each student also notes whether their average was above 2.2 pounds.

Which of the following would you expect to be true about the students' average recorded weights?

a) More of Jaiden's team (20 weighings) will have average weights above 2.2 pounds.

[✓ b)] More of Paulina's team (5 weighings) will have average weights above 2.2 pounds.

c) There is no reason to think that either team's weighings will be more likely to have average weights above 2.2 pounds.

- Small sample size so more likely to have a larger variation in averages)

5.46: S: Okay. So it's geology. In

the geology course, an instructor has her students weigh a metal disc several times on the same scale. The scale is not completely accurate and slightly inconsistent from weighing to weighing. However, the scale is equally likely to read about the true weight, as it is to [00:03:17] read below the true weight. The class is divided into two team led by Jaden and Paulina. Jaden's team decides to weigh the disc twenty times. They compute and record the averages of the twenty Wayans -- weighings. Paullina's team decides to weigh the disc five times, then compute and record the averages of the five weighings. Suppose the true weight of the disk is two pounds. All the students experience -- all of [00:03:43] of the students are experienced with -- with using the scale and record the average weight that they found. Each student also notes whether their average was above 2.2 pounds. Um, which of the following would you expect to be true of the students average recorded weights? More of Jaden -- Jaden's team will have average weights of above 2.2 pounds, more of Paulina's team will have the [00:04:08] average weights of 2.2 pounds, and then there's no reason to think that either team's weighings will be more likely to have an average -- average weights above 2.2 pounds.

S: Um, so... /{from there to there, obviously?}/. Um, I would say that probably Paulina's team would be more likely to have, um, their weights above 2.2 pounds, just because, um, if they have less of a sample, um, size than -- there's more likely like that one weighing will have -- will deviate [00:04:52], like, from that two [OK] pound, um, so it'd gives more of, like, effect. Like, if they have one at, like -- for some reason like -- gave, like, a two point five, and then they also, like -- and their -- keep on like staying more above two than under two, [OK] um, for those like next five times, then if their average is definitely going to be higher, but if they have like 20 times, like, they might have five in a row that are above 2.2, [00:05:17] but they might also have five below 2.2. [OK] So, five doesn't really give like a good, um, mixture. Like, it d-- it just isn't a big enough sample size, so.

I: Okay. OK. So you can indicate your answer, and then write a couple bullet points about why you think that's the correct answer. Okay and how confident are you in your response.

S: I'd say confident.

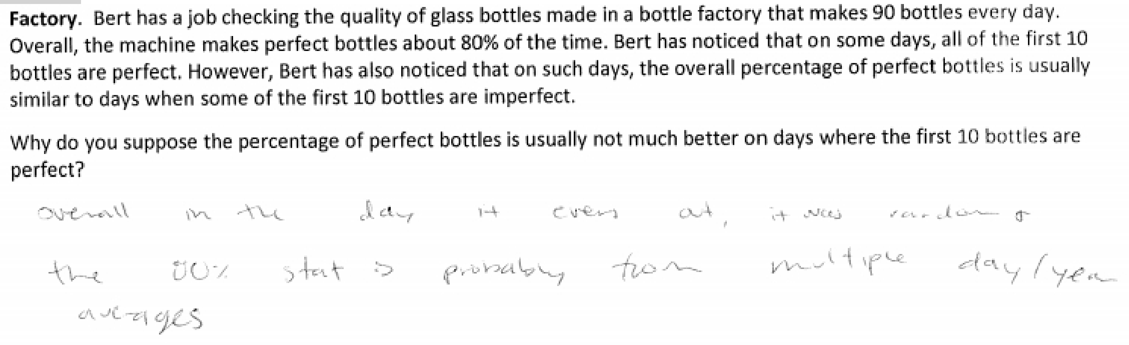
I: Okay. And can you repeat back to me the written problem?

S: So it was in a geology class, so Jayden and [00:06:16] Paulina had two teams, and so Jaden's team wanted to have -- was taking, um, 20 weighings of this metal, um, disc on a scale. And the scale could either go, like, above a certain point, or below it wasn't exactly accurate. Um, and then Paulina's team decides do five trials, like, to see how much, like they thought that it weighed, and took the average. Um, and then it asked which would be more likely to have, like, an average of, like, over a two point [00:06:44] two. And so I said on Paulina's team is more likely, just because of the sample size.

I: OK, great. So, um, /gonna have you/ read the next problem now.

5.128: Bottles (from videorecording)

5.129:



Factory. Bert has a job checking the quality of glass bottles made in a bottle factory that makes 90 bottles every day. Overall, the machine makes perfect bottles about 80% of the time. Bert has noticed that on some days, all of the first 10 bottles are perfect. However, Bert has also noticed that on such days, the overall percentage of perfect bottles is usually similar to days when some of the first 10 bottles are imperfect.

Why do you suppose the percentage of perfect bottles is usually not much better on days where the first 10 bottles are perfect?

overall in the day it evens out, it was random & the 80% stat is probably from multiple day/year averages

5.130: S: OK.
5.131: Burt has a job checking the
5.132: quality of glass bottles made in a
5.133: /inaudible/ -- a bottle factory that makes 90
5.134: bottles every day. Overall the machine
5.135: makes perfect bottles around 80 percent
5.136: of the time. Burt has noticed that on
5.137: some days, all the first ten bottles are
5.138: perfect.
5.139: [00:07:13] However, Burt has also noticed that on such
5.140: days, the overall percentage of perfect
5.141: bottles is usually similar to days when
5.142: some of the first ten bottles are
5.143: imperfect.
5.144: Why do you suppose the percentage of
5.145: perfect bottles is not usually m-- is
5.146: usually not much better on days where the
5.147: first ten bottles are perfect?
5.148:
5.149: S: Um, really
5.150: it's just like i-- like -- probably, um, just, like, chance, like
5.151: [00:07:38] it doesn't really -- it's like a -- if you have like --
5.152: if you're making 90 bottles every single
5.153: day, the first like 10 won't really make
5.154: an effect. Like, you have to look at the
5.155: overall, because anything could happen right in
5.156: the beginning [Hm], like you could have ten
5.157: perfect bottles, or ten imperfect bottles,
5.158: or some kind of mixture, but overall, um, like,

if they make about 80 percent of the time, I'm sure that they had the -- they got [00:08:00] that percentage over, like, quite a few days. [Hm] Like, even maybe years. Um, and so you really have to look at like the overall day rather than the first ten, cuz like, it's kind of like with the blocks how you could, like, have two white blocks in the beginning, and you have two black box -- black bo-- blocks in the beginning, but it's it-- still the same percentage of getting like two black [00:08:22] blocks in the beginning, and then two white blocks at the end, so.

I: Okay. Um, so I didn't quite follow what you just said about the -- the blocks. Could you tell me a little bit =more about that?=

S: =Yeah, so like= when you draw, say, four blocks, you could have two blocks -- you could draw two white first, and then you could draw two black. [OK] So then, it's a 50/50, like you have -- like out of the four blocks that you got, you got two white, two black. But then you could [00:08:48] also, like, if you did another round, you could have got like, say, two blacks, and then two whites, [Hm] and it's still like overall the same percentage, it just depends on like the -- um, like the -- like the, um, way that you grabbed them. Like it's -- overall, like, at the end -- it didn't -- it doesn't really matter, you still have like those -- that, like, the same amount of, like, white and black. [OK] So it just kind of -- but like the order of them [00:09:12] changed, so. [OK]

I: And could could you relate that to this problem, what you were just saying?

S: Yeah, so it's like when he -- he noticed that like some days, like, the ten bot-- the first ten bottles are perfect, and then some days, like, they weren't, like, perfect, like for first ten. So it -- like, but overall, they were still like at that -- like, about 80%. Um, so it really doesn't matter, like -- he -- he could have got like ten perfect [00:09:39] bottles in one day, but then it would ev-- evened out throughout the day, um, just as you make more, so.

I: OK, great, thank you.

Bottles (screen recording starts here)

I: I'm just gonna move this to the side, here, just so we /can see?/. [00:00:11] [Yeah] /cs/ Okay. And how confident are you in your response?

S: I'd say confident.

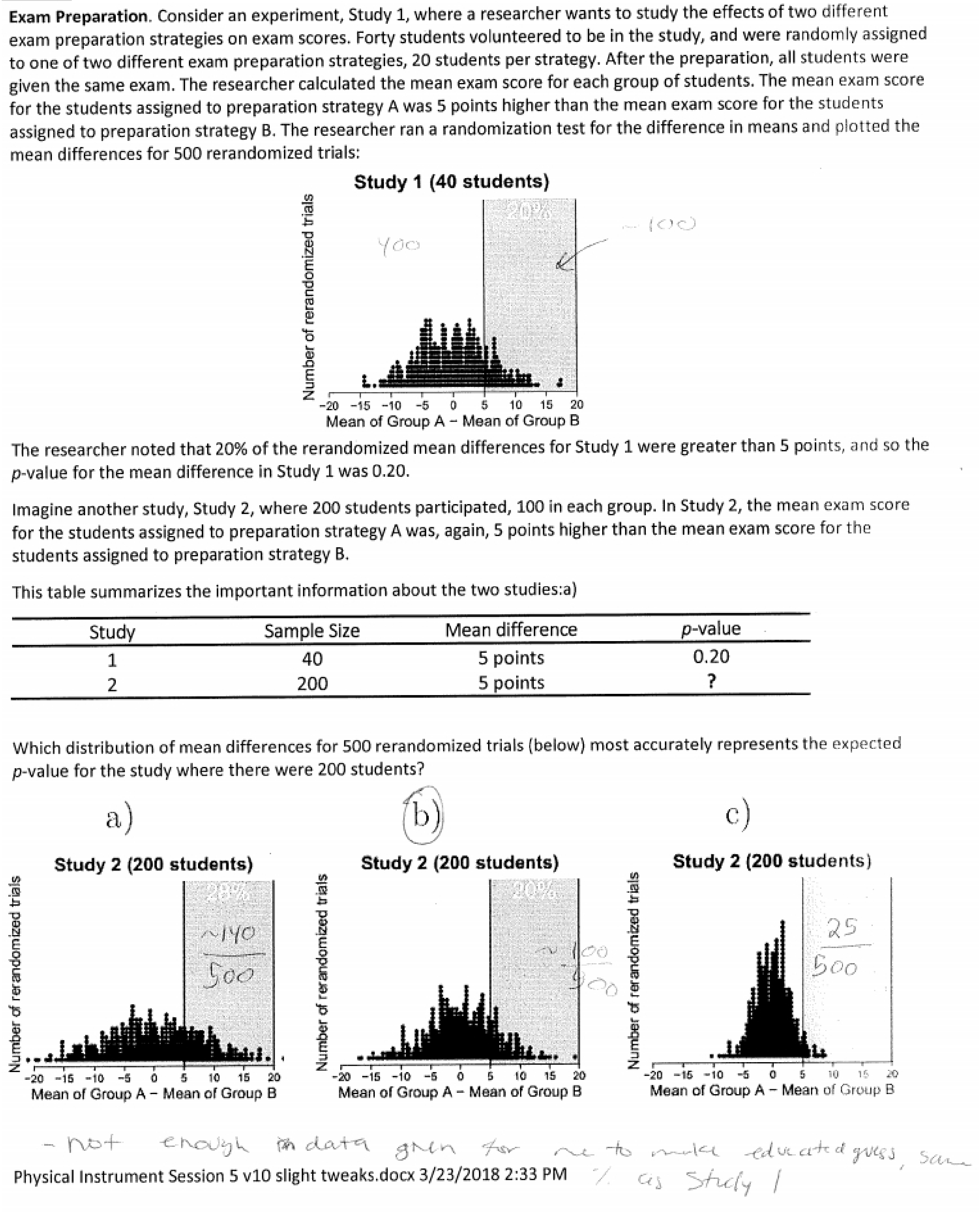
I: Okay. And could you repeat back to me the written problem?

S: Yeah, so it was at a bottling factory and so each day they make around like 90 bottles, and it's about 80% of the bottles that they do make, um, they can use and sell it. They're, like, perfect. Um, and he no-- and one of the workers noticed [00:00:41] that, uh, like some days like te-- the first 10 bottles are perfect, and he was like looking at to see if the -- like the days that were like the first 10 bottles were imperfect were like the same

5.239: percentage and he found that like the
5.240: percentage of like bottles that were
5.241: okay, um, were about the same for each day,
5.242: so then my response to whether like this
5.243: is seemed like correct or some--
5.244: [00:01:06] whatever the -- I can't remember the exact
5.245: wording of it, [mh] but, um --
5.246:
5.247: I: So I just needed the
5.248: problem. [OK] So you don't need to
5.249: summarize your answer. Okay.

5.250: Exam Preparation

5.251:



Exam Preparation. Consider an experiment, Study 1, where a researcher wants to study the effects of two different exam preparation strategies on exam scores. Forty students volunteered to be in the study, and were randomly assigned to one of two different exam preparation strategies, 20 students per strategy. After the preparation, all students were given the same exam. The researcher calculated the mean exam score for each group of students. The mean exam score for the students assigned to preparation strategy A was 5 points higher than the mean exam score for the students assigned to preparation strategy B. The researcher ran a randomization test for the difference in means and plotted the mean differences for 500 rerandomized trials:

Study 1 (40 students)
20%
400
~100
Number of rerandomized trials
−20 −15 −10 −5 0 5 10 15 20
Mean of Group A − Mean of Group B

The researcher noted that 20% of the rerandomized mean differences for Study 1 were greater than 5 points, and so the *p*-value for the mean difference in Study 1 was 0.20.

Imagine another study, Study 2, where 200 students participated, 100 in each group. In Study 2, the mean exam score for the students assigned to preparation strategy A was, again, 5 points higher than the mean exam score for the students assigned to preparation strategy B.

This table summarizes the important information about the two studies:a)

Study	Sample Size	Mean difference	*p*-value
1	40	5 points	0.20
2	200	5 points	?

Which distribution of mean differences for 500 rerandomized trials (below) most accurately represents the expected *p*-value for the study where there were 200 students?

a) [✓ b)] c)

a)
Study 2 (200 students)
28%
~140
500
Number of rerandomized trials
−20 −15 −10 −5 0 5 10 15 20
Mean of Group A − Mean of Group B
b)
Study 2 (200 students)
20%
~100
500
Number of rerandomized trials
−20 −15 −10 −5 0 5 10 15 20
Mean of Group A − Mean of Group B
c)
Study 2 (200 students)
25
500
Number of rerandomized trials
−20 −15 −10 −5 0 5 10 15 20
Mean of Group A − Mean of Group B

- not enough data given for me to make educated guess, same % as Study 1

Physical Instrument Session 5 v10 slight tweaks.docx 3/23/2018 2:33 PM

S: Okay. So, exam preparation. Consider an experiment, study one, where a researcher wants to say the effect of two different exam preparation strategies on exca-- on exam scores. Forty students volunteered to be in the study, and were randomly assigned to one [00:01:32] of two different exam preparation strategies, 20 students per strategy. After the preparation, all the -- all students were given the same exam. The researcher calculated the mean sc-- exam score for each s-- group of students. The mean exam score for students assigned to preparation study A was 5 points higher than the mean exam score for the students assigned to preparation study strategy B. [00:01:58] The researcher ran a randomization test for the difference and means part of the mean differences for 500 rerandomised trials. Um, the researcher noted that 20% of the rerandomized mean differences for study one were greater than five points, and so the p-value for the mean difference in study one was zero point two zero. Imagine an the-- another study, study two, where 200 students [00:02:24] participated, 100 in each group. In study 2 the mean exam score for the students assigned to the preparation study A was again five points higher than the mean exam score for the students assigned to preparation study B. This table summarizes the important information about the two studies. Um, which distribution of mean differences for 500 rerandomized trials, below, most accular-- accurately [00:02:50] represents the expected p-- p-value for the study where there were 200 students.

S: So...

{That's 20 percent.} Um... study two, {two hundred participated}. {5 points higher than the mean exam score.} Um, so I would probably say that -- it would probably, um, be the 20% again [OK], um, just because of the -- the mean was once again, like, the difference in the [00:04:27] means I should say -- was over five. [OK] There was twenty percent up here, so, um, and it was the same -- or I guess that's with forty students. Um, but I guess that's the same percentage, [OK] I would guess. It doesn't really give you that much information, so I would probably say like the -- probably... hmm.

I: Do have any questions about what something means or represents?

S: I mean the way I've usually -- has calculated p-values is, like, [00:05:09] actually having the like how many like numbers are like above a certain like -- most extreme, and you take like a certain point [Hm]. And you don't really have that, so like that's why I don't really know how to get the p-value with just like basically given like only like the mean difference for it, and then like a sample size, like. [OK] But -- I've never calculated like this, so...

I: Okay. So I think this is [00:05:44] calculated the same way that you've been doing in class.

S: We usually have the numbers. Like how many this, and we do that, and then [OK] divided by the sample size. So we don't usually -- we don't work with as many percents I would say. [OK] It's more of like the straight number.

I: Okay. [Yeah] So this is -- these are all 500 rerandomised trials, so, um -- you know, so this would be about, um, so 20%... I mean I could calculate [00:06:16] what those numbers are. Um, so this is calculated that same way, the calculation has just been done for you already. So this was -- in this graph, um, there would have been... times point two -- about a hundred. So there's about a hundred rerandomized trials here. And then 400 here. And similar for this one, because this is the same percentage [00:06:52]. 100, um -- so 100 over 500 would give you that 20%. And then -- so this would be about 140 over 500, over the 500 trials. And this would be, um, about twenty-five of the 500 trials. And so, um, the calculation is the same in each one, and they're all at five and above. But, um, they differ and how the rerandomized trials look. Um, so this, um -- this p value is calculated all the same way, but there's [00:07:42] different numbers of trials that are there or higher in these three plots. So does that, uh, change your thinking any or help your thinking any?

S: No, not really. /c/

I: Okay. All right. /c/ OK. So tell me a little bit more about that. So you're saying in class, um, you usually are not working with percentages as much?

S: Yeah, and especially cuz like if I'm only given like what the answers are potentially, like, I have no way of like [00:08:16] calculating that myself. Like the percentages it sh-- like should be. [OK] So like usually it's like -- okay we have like the graph, and then we know, um, we have like that benchmark. So it'd be, like, the mean difference of like five, and then we would have that like count [Hm] of like how many are more extreme or like equal to that. [mh] And then we would do that over the, um, like number, the sample size.

I: [00:08:45] You would do what over the sample size?

S: The -- the not -- like, the number of extreme values. So all of the -- [OK] like each individual dot that's like above or like in that bracket.

I: Okay. And so that's exactly the same way that this is [Yeah] calculated. Yeah. So there's about a hundred and forty extreme values here out of the five hundred. There's about a hundred [00:09:08] extreme values out of the five hundred, and there's about twenty five extreme values out of the five hundred in this plot. And these are all different scenarios of what could happen, and the question is which one seems most like what you think will happen. I'm just wondering if there's any other information I can give you that will kind of help --um, help your th-- help your thinking about this problem, perfectly happy to do that.

S: I feel like [00:09:50] -- it just -- like I just don't under -- I feel like I just like personally like there's not enough information given [OK] for me to even like give like an educated guess, because [OK] like you don't have any -- like you're basically told like -- okay, so like for this one, it's like okay so you have a sample size of 200, you did the 500 randomized trials, and your mean difference was 5 but you don't have any [00:10:14] other information besides that [/mh?/]. Besides -- the only other information you're given is there was a different sample size, and this was -- these were the results but you don't know, like, what really like -- what the ex-- results were for them, so I feel like there's no way to like accurately say like yeah this is probably what the p-value could be, cuz, like that -- that's just like what I'm thinking.

I: Mh. Okay.

S: [00:10:46] But I mean, I can't -- I feel like with the little information that is given the best answer I would say would be 20%, just because it's the same as study one [OK], and you'll -- almost the same, but it's a different sample size, so like -- so I would just say B just because I really don't know, and I just don't feel like there's enough information for me to even get like an educated guess. [OK] So. [OK] Yeah.

I: Sound good. Um, [00:11:20] and could you write a couple bullet points, about why that's your answer. Okay. And, uh, and how confident are you in your answer?

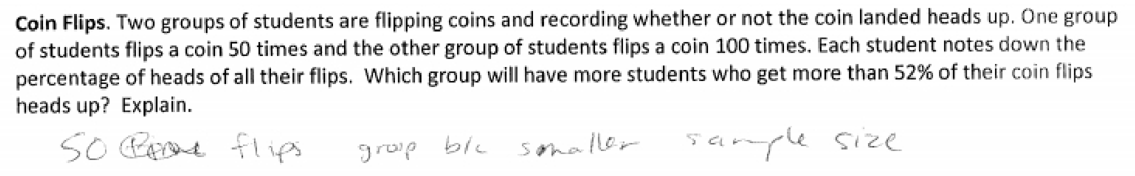
S: Not confident at all. /cs/

I: Okay. OK. And can you just summarize the problem?

S: Yeah. It was about two studies with, um, two different exams. And one of the studies had 40 sample-- or 40 kids, and then the other had f-- um, 200. And it was -- ask -- it gave you like the p-value, and it was asking, um, what the p-value was for [00:12:19] the 200 randomized study.

I: Okay.

Coin Flips



Coin Flips. Two groups of students are flipping coins and recording whether or not the coin landed heads up. One group of students flips a coin 50 times and the other group of students flips a coin 100 times. Each student notes down the percentage of heads of all their flips. Which group will have more students who get more than 52% of their coin flips heads up? Explain.

50 flips group b/c smaller sample size

S: Okay. Coin flips. Two groups of students are flipping coins and recording whether or not the coin landed heads up. One group of students flips a coin 50 times, and the other group of students flips a coin 100 times. Each student notes down the percentage of heads on -- of all of their flips. Which group will have more students who get more than 52 percent of [00:12:48] their coins flip up -- up -- on -- or flip head -- h-- heads up?

S: Um, I don't really think you can -- I would say probably the one group with -- that's only slip-- flipping it 50 times, but I think with each group, um, you're gonna get more variation in like the 50 times, just because you don't have as many trials, so like one trial -- like one, um, head or tail will definitely change the percentage [00:13:19] more, just because the sample size is lower, so like the proportion of it. Um, but it could go up or down. It won't just go up. But, um, like I think -- especially because 52% isn't that much more than 50% [OK] which, um, so I feel like they probably will be -- it could be either one, but to give like a one answer or the other, I would say just the -- the 50 times coin flip group, just because there's less of [00:13:49] a sample size, so it would deviate more. But it could go either above or below 50.

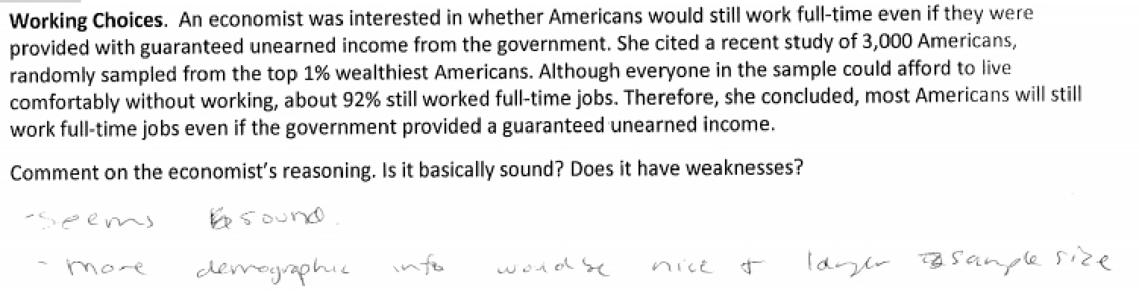
I: Okay. Great. And how confident are you in your response?

S: I would say confident.

I: Okay. And could you summarize the written problem?

S: Yeah. So it was asking about coin flips. So one group had-- was gonna flip their coin 50 times, and then the other was going to flip their coin, um, 100 times, and recording [00:14:37] the amount of heads that they got. And it was asking which one would be above 50 -- more likely to be like above 52% for heads. [OK]

Working Choices



Working Choices. An economist was interested in whether Americans would still work full-time even if they were provided with guaranteed unearned income from the government. She cited a recent study of 3,000 Americans, randomly sampled from the top 1% wealthiest Americans. Although everyone in the sample could afford to live comfortably without working, about 92% still worked full-time jobs. Therefore, she concluded, most Americans will still work full-time jobs even if the government provided a guaranteed unearned income.

Comment on the economist's reasoning. Is it basically sound? Does it have weaknesses?

-seems sound.

- more demographic info would be nice + larger sample size

S: Working choices. An econo-- economist was interested in whether Americans would still work full-time, even if they were provided with guaranteed unearned income from the government. She cited a recent study that -- of 3,000 Americans [00:15:06] randomly sampled from the top 1% wealthiest Americans. Although everyone in the sample could afford to live comfortably without working, about 92 percent still worked full-time jobs. Therefore, she concluded, most Americans will still work full-time jobs even if the government provided a guaranteed unearned income. Comment on the economist reasoning. Um, is it basically sound? Does [00:15:31] it have weaknesses?

S: OK, so I would say just looking at it, it seems like it would be sound. Um, 3000 isn't a ton of Americans, but then again there's not that many like wealthy like one -- out of like the one-percent wealthiest Americans. I'm sure that proportion is quite high [OK], um, out of all of them. Um, so I feel like that probably is fine. I'm not sure how many -- how much is like the 1% [00:16:03] wealthiest [mh]. It could be bigger, but I feel like 3,000 as a decent amount. [OK] Um, 92% at working full-time jobs, this seems like fine to me. I know I work at -- one of it -- my jobs I know, like, some of the people that go there, like, have a lot of money, and -- but they also spend a lot, and so like it makes sense that people would still want to, like, work because they're spending almost everything so it's like they may [Hm] [00:16:34] seem wealthy, and they have a lot, like, s-- a lot of things that cost a lot of money. [Hm] So it'd still, like, put their wealth up there, but they don't actually have a lot, like, actual money. Um, and also I feel like if you have that much money, you probably like what you do, [Hm] and so I feel like if you really enjoy it you do, like, I wouldn't want to stop really working. And I'm sure some of the really wealthy [00:16:58] Americans, you know they could be a pop artist or something like, and singing is what they love to do so it's like why would I stop singing just because like I have enough money, and for them I'm sure the money part really doesn't mean anything. It's more of what they really like to do. And -- and some people just really love working. /c/ [mh] So I feel like it makes sense to me, and I think it seems [00:17:23] sound. If anything, a bigger sample size would have worked, but I'm not sure of the percentage of the 1% wealthiest Americans, and having 3,000 or -- like what also like you don't really -- she doesn't really explain in this sh-- like, short little snippet, like, what the demographics, like, what part of the area they live in as [mh] al-- also, or stuff like that, just like the smaller, like, more of just [00:17:46] like who they are as a person, like what's their background, so.

I: Okay. And so how would the -- how would -- if you did know some of that demographics, how would that affect your decision?

S: Um, I think, i-- looking at like what part of the country they live in, and probably what they -- how they grew up, like looking at like where they grew up, [Hm] what their background is. Because that can give a huge impact on like what they do [00:18:19] and like they what they want to do. [Hm] Like if they -- and also like what kind of job they have, I mean if they're somebody that works in the business field, they could potentially just be working cuz they like to work, they don't -- may not -- they don't probably -- I mean they d-- may not need to work, but I know like with like say like a football player, like they're gonna play as long as they can [Hm] and with [00:18:42] like sports and stuff, and they also make a lot of money, but really the only reason they quit is because they get old. [Hm] So, um, I would say just like looking at what their occupation is and kind of maybe putting them into groups [OK] and look-- looking at it -- you can look at it h-- like definitely looking at it holistically as well, but making sure that you have a variety of [00:19:05] different, um, career paths to get [OK] a more accurate definition, so.

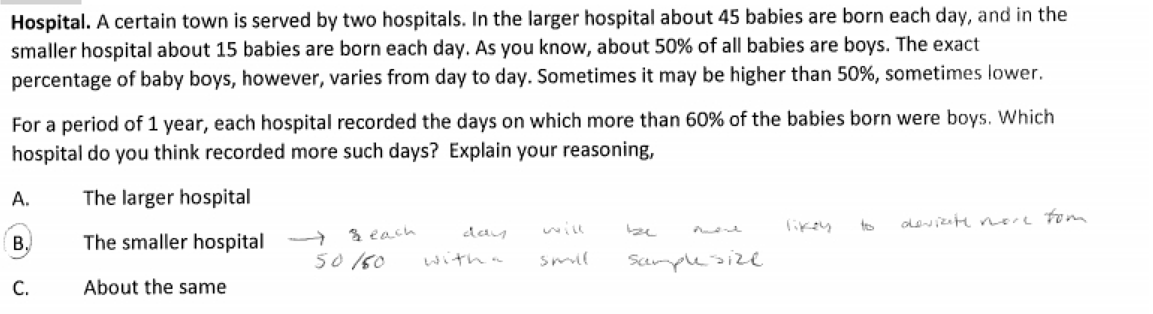
I: So you'd want to see that there was a variety of career paths and in order to kind of trust the study. [Yeah] Okay. Great.

S: Should I write that down? [Yeah] OK.

Pre-questions introduction

I: OK. So now I'd like to return to the problems we did on the first day. Um, so again please read each problem ar-- aloud, and think aloud as you saw each problem. [OK]

Hospital



Hospital. A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50% of all babies are boys. The exact percentage of baby boys, however, varies from day to day. Sometimes it may be higher than 50%, sometimes lower.

For a period of 1 year, each hospital recorded the days on which more than 60% of the babies born were boys. Which hospital do you think recorded more such days? Explain your reasoning,

A. The larger hospital

[✓ B.] The smaller hospital → ~~3~~ each day will be more likely to deviate more from 50/60 with a small sample size

C. About the same

S:

[00:20:06] Okay. So, hospital. A certain town is served by two hospitals. In the larger hospital about forty-five babies are born each day. And in the smaller hospital, about 15 babies are born each day. As you know, about 50% of all babies are boys. The exact percentage of baby boys, however, varies from day to day. Sometimes it may be higher than 50%, sometimes lower. For a period of one year, [00:20:33] each hospital recorded the days on which more than 60% of the babies born were boys. Which hospital do you think, um, recorded more such days? Explain your reasoning.

S: Um, I would say the smaller hospital, just because it has a, um, much smaller sample size. So one baby boy is

5.669: gonna to change the percentage and like
5.670: the proportion of, um, the babies born each
5.671: day quite drastically, compared to say 45
5.672: [00:20:59] babies. [OK] Um, so I -- I would probably say the hos--
5.673: the smaller hospital for that
5.674: reason.
5.675:
5.676: I: Okay. So you can indicate that, and write a couple bullet points about your
5.677: reasoning.
5.678:
5.679:
5.680: S: I guess over...
5.681: Actually, I'm going to change my answer. Uh, it's --
5.682: because I forgot that was over a period
5.683: of one year. [OK] Um, I would probably say
5.684: that they're about the same. [OK] Yeah.
5.685: I would say p-- m-- {which hospital do you think recorded more such days...} Uh. No. I'm going to stay with my -- smaller hospital.
5.686: [00:21:41] [OK] Yeah. Cuz it's like -- I j-- I going
5.687: to change it, but then when I read, like,
5.688: just which hospital do you think recorded
5.689: more such days, so overall I think
5.690: they would probably gonna be around 50 each,
5.691: just because 365 days [Hm] -- that's a lot of
5.692: days that you can get data for. [mh] But
5.693: looking at it individually, and which
5.694: days, like -- you know they could have 10 --
5.695: like they could have like seven baby
5.696: [00:22:08] girls or something. [mh] Or like eight -- or
5.697: like ten baby girls. Um, and that's
5.698: definitely gonna, like, increase or
5.699: decrease the percentage. [mh] And -- but overall,
5.700: like, they could also have like the same
5.701: for boys. So I would say still the
5.702: smaller hospital, just because it's more
5.703: likely to deviate, [OK] um, while like the -- the bigger
5.704: hospital probably has more of like -- it
5.705: doesn't go up and down quite as
5.706: [00:22:32] drastically. [OK] So.
5.707:
5.708: I: And so can you tell me

a little bit about -- um, you said something about -- something about the going for what period of one year sort of was leading you in a different direction for a moment. [Yeah] Can you tell me about that?

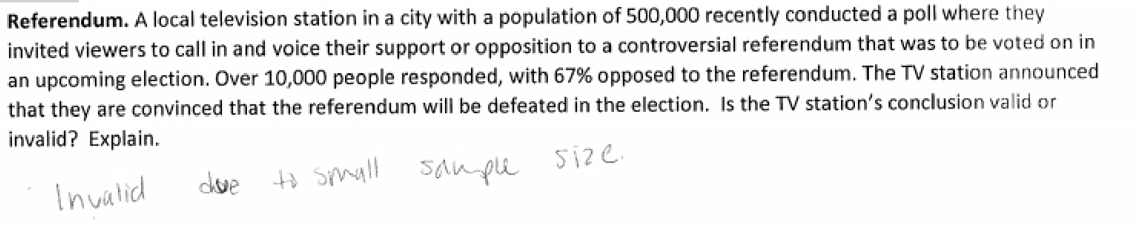
S: So I was thinking like over a year, they're gonna be about the same, like -- it's kinda like flipping a coin, like, if you only do it like say 15 times, it's gonna -- you'd have no idea what your percentage is [00:22:57] gonna be [mh] just because like it could really vary. But if you do it a hundred times, it's not gonna vary as much, it's probably gonna be closer to 50%. Um, looking at it, like, over a year, um, compared to like one or two days, or, say, like a week, um, it's probably then -- gonna be around that 50%. [OK] So.

I: Okay. Great. So can you put a few bullet points? {So I'll put this over here.} [OK] Don't need it yet but just... And, uh, how [00:24:07] confident are you in your response?

S: I'd say confident.

I: Okay. I won't make you [/c/] repeat them back to me on these ones.

Referendum



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

· Invalid due to small sample size.

S: Okay. Um, referendum. A local television station in a city with a population of 500,000 recently conducted a poll where they interview-- invited viewers to call in and voice their support or opposition to a controversial referendum that was to be voted on in an upcoming election. Over [00:24:38] 10,000 people responded, with 67 pr-- opposed to the referendum. The TV station announced that they were convinced that the referendum will be defeated in the, um, election. Is the TV station's conclusion valid or invalid? Explain.

S: I would probably say invalid just because you really never know, um, with, like, elections, especially. But the sample size they took out of the population of 500,000 [00:25:10] depending on where they went in the city, also, like they could have gone into an area that's much more like towards one side of like -- and -- like a political area, like even, like, you know counties, some areas are like a lot more like say liberal rather than conservative and whatnot. So, it doesn't really -- like if they got, if they only really like had one section of the city call in, and they didn't really [00:25:36] look at that, um, it can definitely be, like, mm-- definitely swayed over one way, and 10,000 ap-- out of 500,000, like, that's very small. [OK] Um, and you know, some people -- like the people that are op-- very opposed to the referendum, um, they could be very vocal people rather than [Hm] the people that actually care, like actually would, um, want the referendum to pass. Like, they [00:26:03] could be more of like quiet and not really like wanting to call in. [mh] Um, I just don't think that, like with -- if it was a bigger sample size, um, you could overlook those things I think. [Hm] But, having a sample size of -- not even a -- like a quarter of the population, it's very small. I mean it's a totally [OK] decent size, but like it's still very small overall. [OK]

I: [00:26:46] Um, so you said something about the small sample size, you also said something about people being vocal or not vocal. [Yeah, um...] Can you tell me more about that?

S: Depending on, like, who calls in, like some people -- it could potentially be with this election where the very us-- s-- s-- sometimes there's like one side that's like more vocal or -- there's certain people that are more vocal. And it could be in this case [00:27:08] like where the people that are like -- they want to call in, and like voice their support or opposition for this, like it could be that way, um, where like most of them that are more likely to call in are opposed, or they were given like notifications on like whether or not it -- like to oppose the referendum. Because maybe the ch-- news station also doesn't want the referend-- referendum to pass. [Hm] And [00:27:34] they want to influence people's voting. [Hm] Because some people will vote, like, based on like, oh I don't think this person's gonna win, or I don't think this is gonna pass, so I'm gonna make sure about for like the right -- like [Hm] the thing that's not gonna pass. Like, um, and so I think -- you don't really know if it's like a truly random area. Um, I just -- yeah.

I: And so what could you do to improve the -- improve the poll?

S: I would say [00:28:02] probably making -- like, I'd stay -- still say, like, people can -- and like -- say like yeah, call in but making sure that when you do, like when you're giving -- you're inviting people a -- viewers to call in, um, I would say don't just restrict it to the viewers, because if you are a TV news station, most of the time you're gonna be leaning one way or the other [Hm] and so I think it's really more [00:28:31] important to, um, open it up to the public rather than just the viewers [Hm], because not everybody's gonna listen to a news station. Like, I personally like wouldn't listen to like a news station on the radio, like it's probably one of the last things I would do. [Hm] And so their population is probably like -- like looking at the demographics, too the people that are actually getting [00:28:53] affected by the -- um, like the referendum, they don't -- probably don't know what the news station -- [Hm] and they wouldn't be giving their voice even though they could vote. [OK] So.

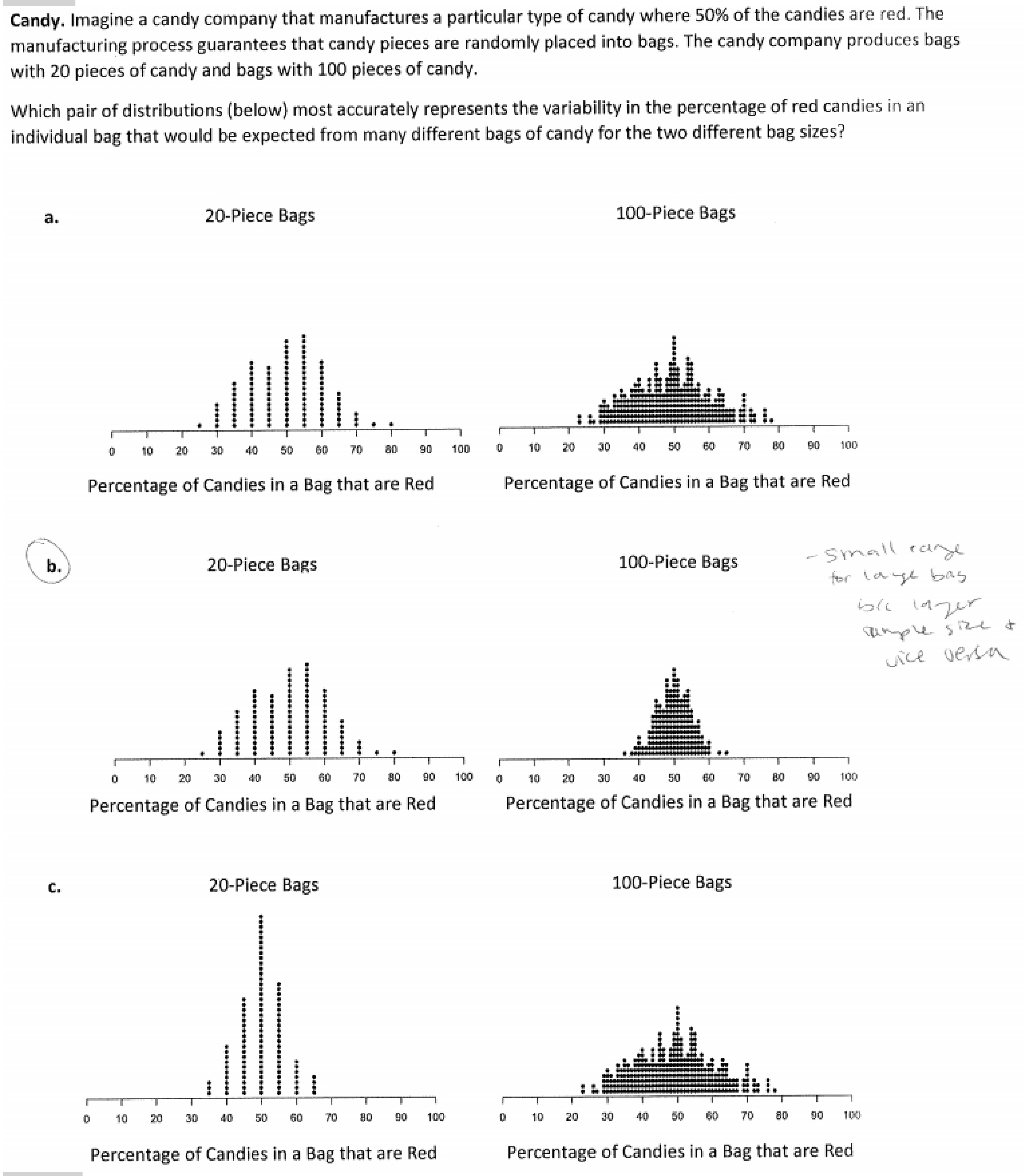
I: OK. Great. Um, and how confident are you in your response?

S: I'd say confident.

I: OK.

5.860: Candy

5.861:



Candy. Imagine a candy company that manufactures a particular type of candy where 50% of the candies are red. The manufacturing process guarantees that candy pieces are randomly placed into bags. The candy company produces bags with 20 pieces of candy and bags with 100 pieces of candy.

Which pair of distributions (below) most accurately represents the variability in the percentage of red candies in an individual bag that would be expected from many different bags of candy for the two different bag sizes?

a. 20-Piece Bags | 100-Piece Bags

Percentage of Candies in a Bag that are Red | Percentage of Candies in a Bag that are Red

[✓ **b.**] 20-Piece Bags | 100-Piece Bags

- small range for large bag b/c larger sample size & vice versa

Percentage of Candies in a Bag that are Red | Percentage of Candies in a Bag that are Red

c. 20-Piece Bags | 100-Piece Bags

Percentage of Candies in a Bag that are Red | Percentage of Candies in a Bag that are Red

5.862: S: Candy. Imagine a candy company
5.863: that manufactures a particular type of
5.864: candy where 50% of the candies are red.

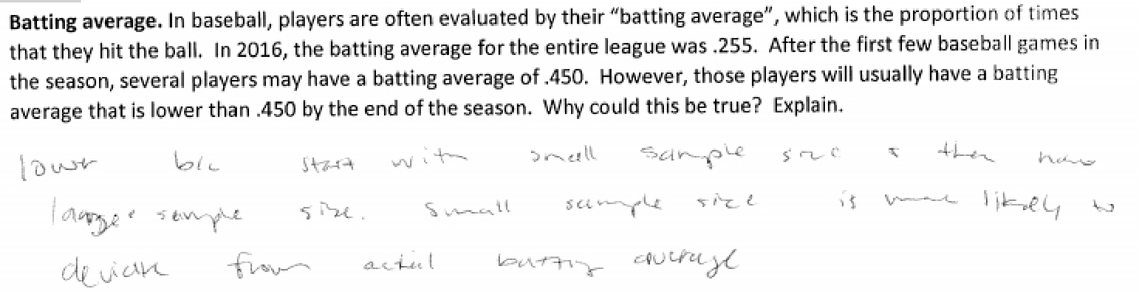
The manufacturing process guarantees that candy pieces are randomly placed into bags. The candy company produces [00:29:28] bags with 20 pieces of candy and bags with te-- 100 pieces of candy. Which pair of distributions below most accurately represents the variability in the percentage of red candies in an individual bag that would ex-- be expected from many different bags of candies for the two different, um, bag sizes. So I would say that it would be... B. Because, kind of like, before if you have a [00:30:02] smaller sample size, one red candy is gonna affect the percentage, or the proportion a lot more, um, and so you're gonna have a much larger range of what it could be. It's like when you get -- like I remember on this day I kinda talked about like fruit snacks. It's like, with the Scooby Doo fruit snacks, say you only have like ten snacks. And like ten little fruit snacks in there, like, some -- I've always [00:30:23] wanted the blue ones. And sometimes there's like not any blue ones, [Hm] and sometimes their four. R-- rather than like -- if there was like a giant fruit snack, you'd be a lot more likely to have like a higher proportion -- or like you'd be more likely to have them, just because like you a lot more options, but it wouldn't -- like it would probably be more of like say like a 50/50 split [mh], um, and so like that's [00:30:48] why it would be the -- the smaller ranges, is that you have less -- each one will like contribute less to the percentage than with the smaller sample size.

I: Okay. Um, can you write a couple [Yeah] bull -- bullets?

5.906:

5.907: **Batting Average**

5.908:



Batting average. In baseball, players are often evaluated by their "batting average", which is the proportion of times that they hit the ball. In 2016, the batting average for the entire league was .255. After the first few baseball games in the season, several players may have a batting average of .450. However, those players will usually have a batting average that is lower than .450 by the end of the season. Why could this be true? Explain.

lower b/c start with small sample size & then have larger sample size. small sample size is more likely to deviate from actual batting average

5.909: S: Batting Average. In baseball player-- in
5.910: baseball, players are often evaluated by
5.911: their batting average, which is the
5.912: proportion of times that they hit the
5.913: ball -- hit the ball. In 2016, the batting
5.914: [00:31:39] average for the entire league was 0.255.
5.915: After the first few baseball games
5.916: in the season, several players may have a
5.917: batting average of 0.5 -- or 0.45. /yawns/ Excuse me.
5.918: However, those players will usually have
5.919: a batting average that is lower than
5.920: 0.45 by the end of the season. Why could
5.921: this be true? Explain.
5.922:
5.923: S: Um, so for this one,
5.924: it's usually -- they might start off the
5.925: season really strong, um, but it's kind of
5.926: [00:32:09] like having that small sample size -- like,
5.927: it's gonna be able to deviate quite a
5.928: bit. Like, you could have -- um, so it was like, it's
5.929: first few bre-- baseball games, you could
5.930: have like a really bad one, and a really
5.931: good one, but it would still probably be
5.932: somewhere in the middle, and the next on
5.933: you have [Hm] would still shoot it up
5.934: quite a bit, because of the small sample
5.935: size. Um, but having, like, you know, multiple
5.936: [00:32:30] games, like, quite a few games. Um, you know

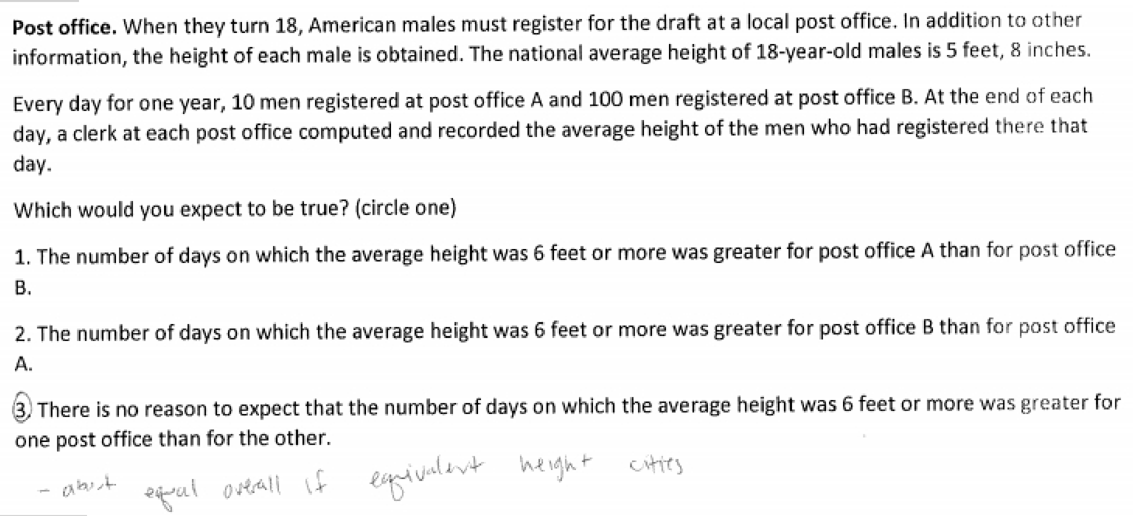
your-- your average is gonna be lower, just because you're looking at so many more. Um, I know, like, with like, if you like score a goal, like if you only look at two or three games, say you scored -- in every single game, you're gonna have a 100% scoring record. Like, you're gonna say you scored in 100% of your games. While, like, if you look at overall, like a season, or [00:32:56] like, even a -- like, two seasons. You're gonna have a much lower -- you're not gonna score every single game most likely. Like, it's -- usually that doesn't happen. [Hm] Just because, you know, depending on who the -- like, you could be playing somebody that's really good, or just -- you're look-- you-- like, you could have off games. Like, you're not gonna be able to be a hundred percent [00:33:16] every single time. Like, you could get injured and miss that game and whatnot. So, um, it's just gonna go down. It's -- like, it'll just be more of, like, every single time you're gonna get some kind of average, and it's gonna be in -- somewhere in the middle between your good and bad, it's not going to be so much more towards like one way or the other. So.

I: Okay. So can you write some of that down? And, um -- [00:34:33] and how confident are you in this response?

S: I'd say confident. [OK]

5.973: Post Office

5.974:



Post office. When they turn 18, American males must register for the draft at a local post office. In addition to other information, the height of each male is obtained. The national average height of 18-year-old males is 5 feet, 8 inches.

Every day for one year, 10 men registered at post office A and 100 men registered at post office B. At the end of each day, a clerk at each post office computed and recorded the average height of the men who had registered there that day.

Which would you expect to be true? (circle one)

1. The number of days on which the average height was 6 feet or more was greater for post office A than for post office B.

2. The number of days on which the average height was 6 feet or more was greater for post office B than for post office A.

[✓ 3.] There is no reason to expect that the number of days on which the average height was 6 feet or more was greater for one post office than for the other.

– about equal overall if equivalent height cities

5.975: S: Post Office. When they turn 18, American
5.976: males must register for the draft at a
5.977: local post office. In addition to other
5.978: information, the height of each male is
5.979: obtained. The national average height of
5.980: 18 year old males is 5 feet 8 inches.
5.981: Every day for one year, 10 men registered
5.982: at a post office A, and 100 men
5.983: [00:35:04] registered at a -- at post office B. At the end
5.984: of each day, a clerk at each post office
5.985: computed and recorded the average height
5.986: of m-- the men who had registered there that
5.987: each day. Um, which would you expect to be
5.988: true? The number of days which the
5.989: average height of 6 feet or more was
5.990: greater for post office A than for
5.991: post office B, the number of days on
5.992: which the average height of -- was 6 feet
5.993: [00:35:28] or more was greater for post office B
5.994: than for post office A, or there's no
5.995: reason to expect that the number of days
5.996: which the average height of 6 feet or
5.997: more was greater for one post office
5.998: than for the other.

S: Um, I would say... that...

overall... {There's no reason to expect the number of days...} I would probably say that overall [00:36:18] there's probably not that much of a difference. Um, but... yeah, I would say that there's probably not that much of a difference, just looking at -- cuz I think each -- each post office, if, say, they're equivalent cities for height [mh], I really don't think that -- is gonna be that much of a difference, um. Yeah. [OK] Because, I think the ten -- the post -- post office A will deviate more, um, but they could also go [00:36:53] much lower, much higher. [OK] And the same with post office B. I don't really think -- I think po-- the 100 -- with the post office B will have -- probably be really close to the 5 feet 8 inches, but it really -- I think it would -- I think they're probably about equal. OK.

I: Um, and so you said something just now about how you'd expect the post office B to have -- something really close to 5 feet 8 [00:37:41] inches. Can you tell me a little bit more about that?

S: Yeah. So if they're saying that, like, um, their average height for their city is probably pretty close to the national average, um, overall they're probably gonna have a higher -- they're probably gonna be more likely to have an average, um, for a year of people that registered that were, uh, 5 8, just because they have a larger sample size, so it's [00:38:05] kind of like flipping -- flipping the coin, like they probably are gonna be much closer to that 5 8 because they've had a larger sample size. Um, while, like only -- having only 10 men, it really, you know, it -- on like say the first day, like, if you have only 9 -- or, like nine that are 5 8 and then one that's like 6 foot or something, um, that's gonna be a 10% difference, that's definitely gonna affect, like, the like [00:38:34] average height and everything like that. Um, so, if, like, you have two men that are over -- like compared to like two out of 10 that are over 6 feet, compared to like 2 out of like 100 that are over 6 feet, um, it's a much smaller percentage ish-- for this the larger sample size than it is for the smaller. [OK] So. I think the -- the ten men just more likely to really go u-- like over [00:39:00] like the national average or under, so it's gonna kinda vary. But over -- I think it'll be close to a five eight, but the hundred men will definitely be like probably pretty spot on.

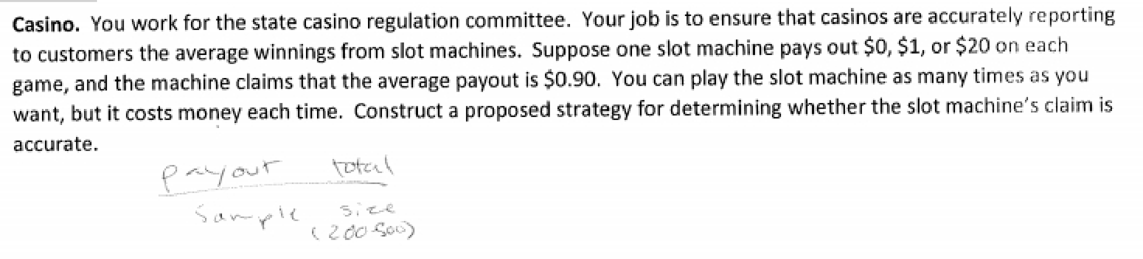
I: Okay. And, um, so -- um, and so with that, why is -- um, why -- why is it about the same? Why are you choosing number three?

S: So with the number of people that are six feet or over, um, the average for -- of height for each post office is gonna be [00:39:36] probably the same [OK], um, just overall. It'll be very close to five eight, but I don't think there's any reason to think, um, that there's gonna be a higher amount of days that, like, the average height was over six feet for one or the other. I think, like, ten men might -- like the post office A might have more numbers, but I feel like over a year, they're probably gonna be about the same overall for, um, [00:40:04] which is gonna be over six feet, so.

I: OK. Um, and how confident are you in your response.

S: I'd say... confident. [OK]

Casino



Casino. You work for the state casino regulation committee. Your job is to ensure that casinos are accurately reporting to customers the average winnings from slot machines. Suppose one slot machine pays out $0, $1, or $20 on each game, and the machine claims that the average payout is $0.90. You can play the slot machine as many times as you want, but it costs money each time. Construct a proposed strategy for determining whether the slot machine's claim is accurate.

$$\frac{\text{payout total}}{\text{sample size}}$$

(200-500)

S: Casino. You work for the state casino regulation committee. Your job is to ensure that casi-- casinos are accurately reporting to customers the average winnings from slot machines. Suppose one slot machine pays out /yawns/ zero [00:40:35] dollars, one dollar, and twenty dollars on each game, and the machine claims that the average payout is ninety cents. You can play the slot machine as many times as you like -- as you want, but it costs money each time. Construct a proposed strategy for determining whether the slot machine's claim is accurate.

S: Um, so I -- what I would do is that either there -- have like a group of us playing the slot [00:40:59] machine, or I could just do it all by myself, but having it -- it prob-- it would go faster if you had like multiple people, like, [mh] playing each slot machine, and, um, I guess -- I -- I guess it's only for one machine, though. So you would have to have just one person doing it. But, um, I'd have quite a few samples, or like quite a few times that you, um, actually played the slot machine. So I -- I would say like at least -- at [00:41:31] minimum 100 [OK] overall. But, um, maybe like whatever -- the avera-- maybe like the average of like one day, or something. Say, like -- they only have like two -- they have like 200 people coming in and playing that single slot machine a day, then I'd probably try to like get it close to that, [OK] just to see like what the average of the day would be, [OK] and then I would just see that and then see what the [00:41:57] average payout was over, um, however many times that it was played.

I: Okay. And so -- so first of all, let's just settle on a number. So what, uh -- what's the kind of number that you -- number of times you'd like to play the game?

S: We'll say 200.

I: Okay. And so why do you pick 200 games?

S: Um, I feel like 100 -- like, I'm guessing, like most the time like a casino probably gets at least 100 people [00:42:23] coming in and playing like a slot -- the slot machine, so I'd say, like, that'd be at minimum, like if you couldn't be able to play that mu-- many times, like that would -- like it'd be preferable to get more than what, like, is expected, just to [OK] like really give a -- a -- a really good number, so. [OK]

I: And so, putting aside how many people play it in a day, um, do you think 200 would be enough [00:42:51] to tell whether the machine is accurate or not?

S: Um, I think so. Um, [OK] it's not -- I mean the more, the better no matter what situation it is. [OK] I think for -- I think -- I -- maybe like 200, I say -- 200, say, I would like to say is a good number, um, but of course like more is better, so like 400 or 500 would like be better than 200, [mh] but I'd say like, yeah, like 200's fine. But 500'd be fine, too. /c/ [OK] [00:43:40]

I: And, uh -- um, why would more -- what would you want the sample size bigger? Like, what would be the advantage of that?

S: Um, just like having a bigger sample size, like, you're just always gonna be able to get that -- that proportion and that -- that average is gonna be more and more accurate, the more you get. Um, because, like, one, like, streak of zero, say, is gonna affect it a lot less than, say, like a hundred s-- hundred times playing, [00:44:06] and then like having a streak of ten, because that means that ten percent of the times you got, like, at least, were. While, like, if you increase that number like 10 out of 500 is gonna be much smaller than 10 out of 100. [mh] So.

I: Okay. Great. Uh -- and, uh, how confident are you in you're response?

S: I'd say confident.

Comparison old-new

*Hospital

I: Okay. Um, so next we're going to look back how you originally answered these problems. Um, [00:44:57] I'll put each in front of you, and so let me know what's changed and what has stayed the same in how you reason about each of the problems.

So this is the original, and this is what you did just now. So what's changed, and what's stayed the same in how you reasoned about this problem?

S: Um, I think for this one I was thinking more of like -- I didn't really think, um, about the amount of days that there was more [00:45:42] than 60%. Because overall the -- the average is probably really close to 50-50 [Hm], so it's gonna be about the same. But if you think about, like, the individual days that they were over 60% [OK], um, I think the smaller hospital would be more. So. Looking at -- I -- I think I just didn't read that question all the way. It's like, you read it but sometimes you miss that like the key word in it, sometimes [Hm], I think. [mh] Um, or just [00:46:08] like you think about it differently after you read back, so. [mh]

I: So, um, this time -- were you thinking -- so if you ignored -- so if -- are you thinking if you just took the total percentage for the whole year, is that what you were kind =of thinking on this one?=

S: =Yeah, so like= [OK] -- say like, however -- whatever 45 times, um, 365 is. Like, um, [/inaudible/] looking at like that number, [OK] compared to like the individual days. Cuz I wrote -- I don't know if -- yeah, I mean, [00:46:38] like for this one it's like -- it's just more like that average will be probably the same for each hospital, about 50/50 for babies, boys and girls. But, um, for the number of recorded days, like, I think the smaller hospital would have more, just because it's got a smaller sample size, so it's gonna be able to deviate up and down from that 50/50 more. [OK]

*Referendum

I: So yeah. Again, how your reasoning has [00:47:12] stayed the same, or changed.

S: Yeah, I think I basically said the same thing for each one. On this one I talked about, like, demographic -- I was -- I know I didn't really like write it, but I know if we talked about demo-- I talked about like the demographic and like what their viewer is. I know I mentioned this time that, um, they get their viewers -- like they could have a very like select amount of [00:47:35] viewers, and, like, I didn't really -- I guess like -- they just all have diff-- like they might have a very individual background compared to like the whole population, and it's also a very small sample size, um, and then the political views stayed the same. So, I'd say my reasoning for both of them was very very similar.

I: Okay.

*Candy

S: OK. The candy one. Um, I said the -- it seems like it seems the same. Um, [00:48:09] I put B for both of them, and I said the small bag, um, one candy impacted the percentage much more than one candy in the hundred bag. And I basically said like the small range, um, for the large bag is gonna be -- it -- the range is gonna be small for the large bag basically because the sample size is so large, so it's not going to deviate quite as much from the -- the actual percent it should be, so. [OK]

*Batting Average

S: [00:48:47] So I talked about the sample size in this one. I think that my reasoning was probably the same. The small sample size, you know, is not go-- it's gonna be able to go up and down quite a bit more fluidly than, um, if it's like a small-- a -- a larger sample size, and for this one I just said, lower. I-- it'll -- like it'll probably be true because the smaller sample size, like -- and then, like, having, like changing to the larger [00:49:13] sample size. Um, the smaller sample size is just more likely to deviate from the actual batting average of the person. So I'd say that my reasoning was the same, I don't really think there's any difference that I approached the problem with. So. [OK]

*Post Office

S: Um, I think this one -- um, I almost, like, kind of corrected myself, I think from the [00:49:45] hospital for this one. Um, I didn't really think about it, probably. But, uh, because I talked about it -- I specifically said the amount of days collected is probably going to be the same, um, so I think that I went -- uh, went by it the same way. Like, the, like, overall like in a -- like a week, the percentages might be drastically different, but if you look at the specific days I feel like it'd probably [00:50:17] be pretty equivalent. So, I think I went about it the same way. [OK]

*Casino

S: OK, so -- I said the -- I had a really small sample size. Um, casinos are really busy, so I feel like 50 times, so like on a specific machine wouldn't be that much [OK]. Um, so I think I went arou-- I went -- I thought about my strategy the same way [mh], but I think I could have definitely increased the amount of -- the -- increased the sample size. Um, [00:50:58] because I don't feel like the 50 times would give a great average [OK]. Now, I don't think it would be, but at the time I did.

I: And so what -- what do you think changed? Like, why do you think you -- why would you prefer 200 now rather than 50?

S: Um, I think just like I thought more about like the actual casino. Like, a lot of people come and go from a casino [mh], and like -- y-- like an individual person might [00:51:25] play the slot machine -- like say if that's the only thing they like to play or something, they could play it like 10 or

20 times, like, just them. [Hm] So, I don't think -- I think that, like, just on a regular day, a single slot machine sees a lot more traffic than just [Hm] 50 times. [mh] So.

I: So you think this would be more -- this would be more true to [mh], um, what -- the real life circumstance of the casino.

S: Yeah. The bigger the sample size the better. /cs/ [OK]

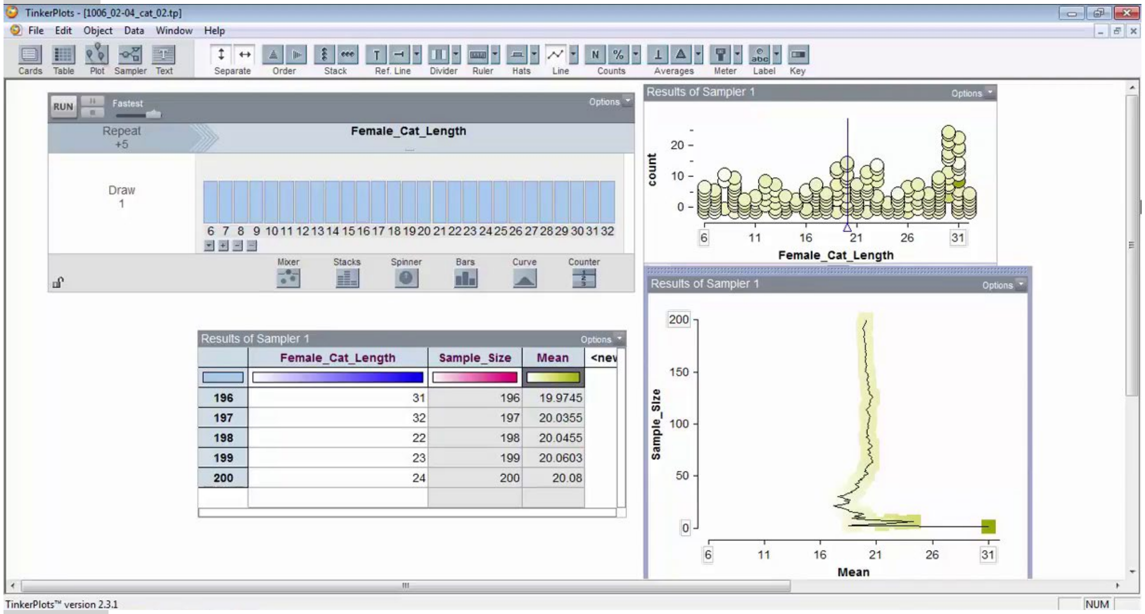
Comparing Representations

I: [00:51:52] So, um, as you know we've looked at both some sample size plots and some combinations over the last, uh, couple weeks. Um, so I wanted to pull up, um, some visuals from activities we did earlier. Uh... /Box folder shows up as empty/ that's weird. /Clicks, out, clicks back in, Box folder now has files/ OK. /c/ [Wow.] I don't know what happened there.

S: Panic.

I: /c/ I was like, there's a lot in this folder. /c/ OK, um, so I'll open up the -- OK -- what we did with the cats. So here's the sample size plot we did. Um, [00:53:14] let's see we can -- I can -- we had this filtered down, but I can make this bigger to the full -- show the full 200 for a moment. Um, so this is the mean and then the sample size, and then the -- how the mean changes as the sample size increases. And, um, I was wondering if you saw -- and you don't need to see anything -- but I was just wondering if you saw any connections between this geology problem and what we were doing

5.1395: [00:53:43] with these plots.
5.1396:
5.1397:

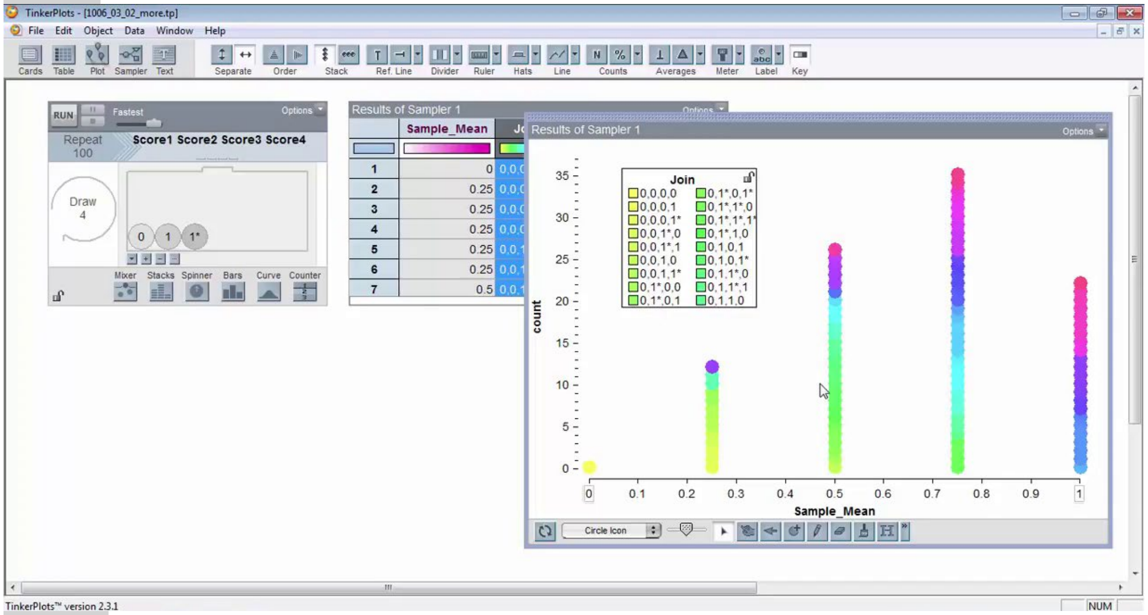


Results of Sampler 1

	Female_Cat_Length	Sample_Size	Mean	<nev
196	31	196	19.9745	
197	32	197	20.0355	
198	22	198	20.0455	
199	23	199	20.0603	
200	24	200	20.08	

5.1398: S: Um. [{Sorry}] No, you're good. Um, I mean it's kind of like -- if you
5.1399: look at -- if the -- the scale... if you think about
5.1400: the scale like kind of in the way of the
5.1401: cat length, um, it's gonna go about even on the
5.1402: same side, so like even that 21 isn't
5.1403: between like the 6 and the 32, um, at one
5.1404: point, like, the average is gonna be
5.1405: somewhere in between whatever two, like
5.1406: ways that it goes. Like, plus or minus, say like,
5.1407: [00:54:23] 5 on like the -- the scale. Um, so it ma-- like, when
5.1408: you look at the graph, um, you know, whatever
5.1409: the -- the range of that -- that scale is
5.1410: probably gonna probably be, the average
5.1411: is gonna be somewhere in the middle, so
5.1412: no matter -- even though like they all have
5.1413: a likelihood of like being the same, um, it's
5.1414: gonna like that average is gonna be in
5.1415: the middle. [OK] So.
5.1416:
5.1417: I: Okay. Um,
5.1418: and we also, um, have looked at, um, these kind of
5.1419: [00:55:03] combination -- these plots that show the

different combinations. Um -- or I can -- yeah. So, do you see any connection between this combination plot and the geology problem?



Results of Sampler 1

	Sample_Mean	Jo
1	0	0,0,0
2	0.25	0,0,0
3	0.25	0,0,0
4	0.25	0,0,0
5	0.25	0,0,1
6	0.25	0,0,1
7	0.5	0,0,1

S: Mm... the only combination I would see is that -- like the -- the order of what they are weighing in, um, if you look at the average, like, you could get like two that are above and [00:55:50] say two that are below, [mh] and it'll give you the same average than, um, like the opposite. So, like, two below and then two above [OK]. Um, that's the only similarity I really see. But, just like there's different combinations to get like that, especially if you do like betweeen weighings, like you could get multiple ones above, and multiple ones below, and then -- or like it could be a different order, [00:56:12] which one like is above the actual average or at -- um, below, so.

I: Okay. And does that -- what

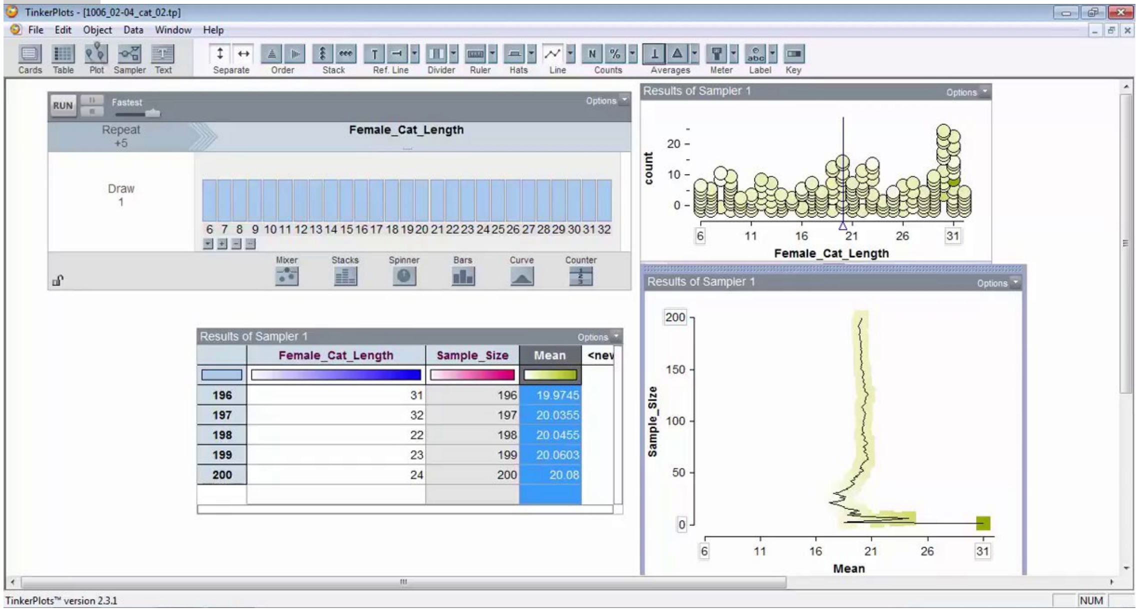
you just mentioned about the order, does that play into anything about how you -- how you would reason about the answer? Does that really to which one of the teams you would choose for the answer?

S: Mm, I don't think so, no. [OK] Yeah.

I: So more of that general similarity.

S: Yeah. I mean, the -- the -- the average isn't gonna change either [00:56:44] way, so. [OK] It's more depen-- for this question I think it's more dependent on the amount of weighings that you do then the -- the combinations, so.

I: Okay. And, um, so we're gonna look at these same two things with the coin flips problem. So again, do you see connections between this coin flip problem and this kind of sample size plot that we were seeing earlier?



TinkerPlots - [1006_02-04_cat_02.tp]
File Edit Object Data Window Help
Cards Table Plot Sampler Text Separate Order Stack Ref. Line Divider Ruler Hats Line Counts Averages Meter Label Key
RUN Fastest Options
Repeat +5
Female_Cat_Length
Draw 1
6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
Mixer Stacks Spinner Bars Curve Counter
Results of Sampler 1 Options
count
20
10
0
6 11 16 21 26 31
Female_Cat_Length
Results of Sampler 1 Options
Sample_Size
200
150
100
50
0
6 11 16 21 26 31
Mean
TinkerPlots™ version 2.3.1
NUM

Results of Sampler 1

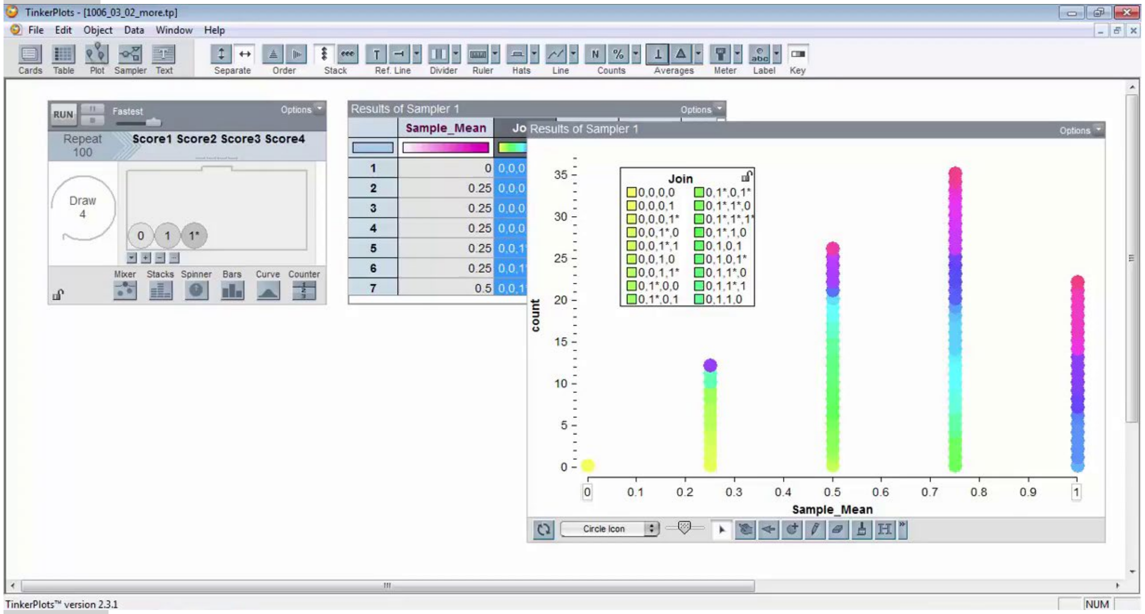
	Female_Cat_Length	Sample_Size	Mean	<nev
196	31	196	19.9745	
197	32	197	20.0355	
198	22	198	20.0455	
199	23	199	20.0603	
200	24	200	20.08	

S: Yeah. Um, I would say more, with, like with the coin flip, um -- you know the average, the mean, [00:57:25] like, say it's like the 50/50 split. It's gonna like what -- the more trials you do it, the more it's gonna really stay at that fifty-fifty overall. Um, while like in the beginning if you only do they say like five flips, it's gonna go up and down, back and forth, quite a bit. Um, it's gonna be quite zigzaggy at the bottom, just like how this one is, but then as soon as you get around like that 50 it [00:57:46] kind of really stays, um, basically almost, like, I would say vertical, and it doesn't really zigzag around. [OK] So.

I: And how does that relate to -- does that relate to your reasoning about the answer to this problem.

S: Um, I would say yes, just because, um -- it -- it's just asking, you know the -- the -- if it's staying, like, around, like you know -- that, like -- like what the percentage is [00:58:13] probably gonna be, um, like the -- like a smaller sample size -- that's more likely to deviate from the 50% so 52 is pretty close to 50, so I'd say like both of them are gonna ge-- probably at one point be 52, [Hm] but the 50 coin flip is definitely gonna be more likely to, like, have that possibility of being over 52, just because like each one is gonna be able to deviate -- like each heads or each [00:58:44] tails is, like -- will definitely affect the -- the percentage of each one quite a bit more than for 100.

5.1511: I: Okay. And again do you see
5.1512: any connections between the kind of --
5.1513: what looking at with these combinations
5.1514: and the coin flips problem.
5.1515:



	Sample_Mean	Jo
1	0	0,0,0
2	0.25	0,0,0
3	0.25	0,0,0
4	0.25	0,0,0
5	0.25	0,0,1
6	0.25	0,0,1
7	0.5	0,0,1

5.1516: S: Yeah, so it's
5.1517: kind of like when you -- you either have like --
5.1518: you can either do heads or tails, um, but,
5.1519: like, if you have a hundred trials, like,
5.1520: you could get a hu-- like 50 tails all in one
5.1521: [00:59:15] row. I mean,
5.1522: it's unlikely, but, and then get 50 heads,
5.1523: or you could also get like 50 heads and
5.1524: like 50 tails but also like 25 and then --
5.1525: like 25 heads and then 25 tails and then
5.1526: 25 head, and
5.1527: twenty-five tails. But the mean will be
5.1528: the same. [OK] Um, so just like looking at
5.1529: the mean, like, the mean is always gonna
5.1530: probably stay around that. Um, there's just
5.1531: [00:59:37] different combinations you can get. But. [OK] Yeah.
5.1532:
5.1533:
5.1534: I: And does that relate to your
5.1535: reasoning about the answer to this

problem?

S: Um, I guess like just a little bit. Not -- not so much, just because I was looking more at the sample size, but different combinations give you the same percentage, um, and would say that's like the only way it would influence my thinking. [OK] so.

Interview closing

I: Okay. Great. So, um, that's all of the kind of [01:00:12] problem-solving. [OK] Um, so we'll have some -- just kind of -- a little bit of just general conversation [mh] about the study. So what do you think these activities and tasks, uh, getting at, as we were [Yeah] -- that we've been doing the last several times?

S: Um, I think a lot of it was probably just how you go around the problem, like how do you approach it, and if you, like, talk about it one day and you talk abou-- and like you do certain activities like [01:00:41] will that influence your -- your final decision, does it make you realize, like does it impact you quite a bit, um, like with your final decision. Um, and also, like, do you become, like -- it's almost like -- yeah, I'd just say, like, the reasoning about how you go around a problem and, um, what not.

I: Okay. And what kinds of -- I mean what -- um, what kind of problem do you think these activities were related to?

S: Um, I would say, like, if [01:01:15] you're like -- for me, I-- whenever I read them it was like -- because like -- it was like I take -- I was taking a test. Like, how would I go about this if I was taking a test [Hm], and -- or like, when I was first introduced to this problem, like how would I go about it if I didn't know, like, what the final answer was supposed to be, like [Hm], um -- and so for me it was like looking at a blank -- blank piece of paper and being like, okay, what does [01:01:39] it -- what first does it want, and then kind of being like what do I know about -- [mh] like from like TinkerPlots the previous assignments I've been doing, so. [mh] Just, like, looking at how like -- which -- what kind of ways that like -- maybe like that most students approach it [Hm]. So.

I: OK. Um, and I was wondering if you -- did you notice anything about your thinking change as we kind of went through the [01:02:07] activities?

S: Um, I'm not sure. I think -- I think my reasoning stayed pretty much the same. I think I changed my -- I would like -- uh -- like if I did one, and then I kind of like looked at the /up?/, I'd be like, oh yeah I re-- like this makes sense, like, [Hm] I remember this in class, like, like, oh wait. I -- I get it now, and then when I did it again, or did something that was similar, I was like -- kind of like was in the back of my mind, [01:02:35] especially because I did the studies like, um, quite -- like, pretty quickly together [Hm] and so, like, it was still fresh in my mind where I could still like remember like what my thinking process was for the first time for some of them. [mh] And like specific anecdotes that I used. Um, [mh] so. [OK] Did I answer the question? [Um...] /c/ I can't remember. /c/

I: Um, so the main -- the main question was w-- what -- did you notice anything change as you went through the study? [Um, I --] In your thinking.

S: Yeah, I don't think [01:03:04] in my thinking it -- I think it was just more of like certain things would remind me of things, or like I would think about it, or like it'd be in class we talked about like the next day, and I was like, oh yeah, like [Hm], =but other than that...=

I: =But what did it --= what's this about -- something came up in class that was related to the study?

S: Well, we're doing p-values right now, so, like [OK] just, like, doing like the [01:03:21] assignments or like, I have -- I have class on Tuesday Thursday, so like sometimes I'd go -- come from class [Hm] on like Tuesday Thu-- if I had -- the study like after one of my Tuesday classes or Thursday classes, it was like TinkerPlots was like fresh in my mind, and stuff. [Hm] But, [Gotcha], I think overall my reasoning, like, especially if I was using TinkerPlots and I'd see the results, it's like okay yeah I get that. Like, [mh] I'm gonna [01:03:44] think about it like that way the next time I approach it, so. [mh] Yeah.

I: Okay. And so if you were gonna explain to somebody, um, why sample size -- why a larger sample size means less variability, or why a smaller sample size means more variability, or -- [mh], um, how would you explain that?

S: I would probably do a

food anecdote. /c/ [OK] Cuz it's like -- like thinking about the fruit snacks, like, I will always remember like in track like [01:04:18] we would always use, like -- we would always get fruit snacks. And if I got the Scooby Doo ones like everybody, like for some reason a lot of our favorite like fruit snack it was, um, like the blue Scooby Snack, [Hm] and, like, we'd always be like oh I hope I get a lot in these and sometimes [Hm] you wouldn't get any, and you'd be so disappointed. [Hm] Um, but if then, it's like if you look in the big picture, like if you [01:04:40] have like a bigger bag, you're gonna be more likely to have that [Hm], or even like if like you have ice cream and you put something in there, and you mix it around. Like you're more likely to have a scoop with like -- if you have like sprinkles or something or like brownie pieces, like the more you put in, the more you're gonna like each have, like in every single scoop rather than [Hm] like some scoops like [01:05:01] may not have any sprinkles or like any pieces of brownie, just because like you don't have as many in there, like [mh], so that would be, like, your sample size. [OK] So.

I: OK. Any more general comments about your participation in the study?

S: Hmm, I don't think so. /c/ [OK] I feel I've answered pretty much -- any -- at least like what what I can think of, if you have any more questions, I'm -- I'm free to -- free -- [OK, sure] feel free to ask them, so.

I: /c/ Sure. Um, what [01:05:28] did you -- what did you like about any of the activities that we did?

S: Um, I liked it was just like -- they were very, like most of them, I would say were pretty straightforward, and it was more of like, you know, like, you don't really have anything to lose with your answer. [Hm!] And so it wasn't like at te-- like, when you're in class, it's like you're taking a te-- if you're taking a test, like, and I would get asked that answer, like, I [01:05:49] would -- I'm, like, more likely to overthink it, I'm like, what are all the variabilities, like, huhhh, I'd be so stressed out, like [Hm] being -- especially, like -- I-- because I'm a physiology major, and so like in my Chem and Physics class, like a certain variability like can really change how you answer like approach a question, even. [Hm] So if you don't have like -- I always like try not to [01:06:10] overthink questions, like [Hm] especially like, I'm like, okay -- they're not supposed -- like for these ones, I'm like they're not supposed to be difficult. Like -- or they're not supposed to be, like -- you're -- you're not like supposed to overthink them, I would say. [Hm] And so I was always trying to like think about, like, okay. Straight forward. Like, /c/ [Hm] like when I was like looking at the hospitals, like, okay, so we [01:06:30] all know that fifty fifty percent for like babies that are born, like they're gonna be boys. But like if you look at like the whole entire population of the world, like girls are actually like m-- like, there's more girls in the world, like [Hm] just because of like how they're -- like they're formed and everything, um [mh], you're more likely to have, like, a girl. So I was like, I'm not gonna think about that [01:06:51] [/c/], like in this, like, it's just not applicable. So I was like always trying to simplify the problem and not overthinking it, [mh] and thinking of more like very straightforward, like -- okay, what is it asking. Okay, think about everything else, but like m-- most likely, like this is what it's probably like really directly asking, so.

I: OK. And, uh, what didn't you like so much about the activities?

S: Um, for some of them I just [01:07:15] felt like I didn't have enough data. [OK] Just, like -- cuz I mean -- I -- I'm like a person that really likes having like a bunch of data before -- like having all the information before I answer questions. [Hm] So sometimes like if it doesn't give like a lot of say, like, numbers or like how st-- like how a s-- like if you give out a percent, like, I'm like, I want to know like how they got their percent. Like [Hmmmm] /c/ [01:07:36] oh, no, no, I'm like the big picture kind of person [Hm], so it's like I want to know everything about like that certain number, and like -- so, for some of them, like, well, I don't -- I'm -- I'm just gonna like go with it, like, I'm gonna say like it's probably accurate, and like just kind of see like where it takes the question and stuff like that, so. [mh] But other than that, like, I don't really think -- I think [01:07:59] it was just, like, pretty straightforward. Like, it's just like questions asking like trying to figure something out, so. [mh]

I: And, uh, can you give an example of one time you felt like you didn't have enough information?

S: I'd say probably like the -- the p-value one. [OK] Just because, like, um, I felt like it was very like -- like for me at least -- like especially cuz I'm not -- we haven't used p-values that often [Hm], but [Hm] [01:08:22] just the way I've always learned to like calculate them is slightly different, and I never had like really approached it that way. [Hm] And so I really didn't know -- I didn't feel like I had like all the numbers I needed to like confidently give an answer. So. [OK] But I'm sure, like, looking at it, for like some people it's like they clicks, but I'm -- also this is -- is like the first stats class I've [01:08:43] ever taken, so [Hm] it's like my mind is like -- I guess I always think about things like slightly different, maybe, [mh] than people if they've taken like maybe one stats class, or they took it in high school, or something, like /c/ [mh]. So.

I: OK. Gotcha. Um, that's all the questions I have. Do you have any questions for me?

S: No. /c/

I: Okay. So you're all set.

S: Perfect.