A Modern Look at Freedman's Box Model

Dennis L. Sun, Joseph Alfredo
California Polytechnic State University

Abstract

This paper revisits the *box model*, a metaphor developed by Freedman, Pisani, and Purves (1978) to explain sampling distributions and statistical inference to introductory statistics students. The basic idea is to represent all random phenomena in terms of drawing tickets at random from a box. In this way, random sampling from a population can be described in the same way as everyday phenomena, like coin tossing and card dealing. For Freedman *et al.* (1978), box models were merely a thought experiment; calculations were still done using normal approximations. In this paper, we propose a modern view of the box model as a practical simulation framework for conducting inference, thus providing a bridge between the box model and simulation-based inference in the education literature. To facilitate this simulation-based approach to teaching box models, we developed an online, open-source "box model simulator" inspired by user-centered design theory to help students scaffold their knowledge. Finally, we suggest some novel ideas for using the box model to teach bootstrapping and discrete probability.

Keywords: box models, sampling distribution, hypothesis test, statistics education, simulation, probability, binomial, hypergeometric, geometric, negative binomial

1. INTRODUCTION

The abstract idea of a "sampling distribution" is challenging for students in introductory statistics courses (Chance, del Mas, and Garfield 2004; Cobb 2007). First, they may confuse the sampling distribution with the distribution of the data (Garfield, Ben-Zvi, Chance, Medina, Roseth, and Zieffler 2008). Second, they have to imagine the sampling distribution from just one realization from that distribution, since the "distribution" is over hypothetical realizations of the data (Cobb 2007). In our experience, even graduate students with several statistics courses under their belt have misconceptions about sampling distributions.

Many strategies have been devised to help students overcome their conceptual difficulties with sampling distributions, including tactile activities and computer simulations (delMas, Garfield, and Chance 1999). One particularly intriguing device for teaching sampling distributions is the "box model," introduced in the celebrated textbook by Freedman *et al.* (1978) (hereafter FPP). FPP were motivated by similar concerns about the difficulty of teaching sampling distributions, as they described later in their instructor's manual (Freedman, Pisani, and Purves 2007):

> One famous difficulty in teaching elementary statistics is getting across the idea that the sample average is a random variable. Randomness, after all, is quite a complicated idea. It is easily overwhelmed, either by the definiteness of the data, or by the arithmetic needed to calculate the average.
>
> In our experience, the most intelligible short explanation goes something like this:
>
> > You took a sample and computed the average. That is a number. But it could have come out a bit differently. In fact, if you did the whole thing all over again, it would come out differently.
>
> This variability is the key point to get across, and it tends to be obscured by the technical sound of the phrase "random variable." As a result, we have given that phrase up—and many other hallmarks of civilization too. For the phrase, at least, there is a good substitute: drawing at random from a box of tickets, where each ticket has a number written on it. This may seem crude, but conveys a clear image.

For the benefit of readers unfamiliar with the box model, the next section summarizes the basic idea. However, we hope that the discussion in the next section will be a refreshing take on the box model, even to those who are already familiar with the analogy.

2. WHAT IS THE BOX MODEL?

A box model consists of a box of tickets, each with a number on it. Draws are made by reaching into the box and drawing tickets "at random." To specify a box model, we first have to say what tickets go in the box. Then, we have to say how the tickets will be drawn: the number of tickets to draw and whether the draws will be made with or without replacement. Finally, we calculate some summary statistic of the numbers that were drawn, such as the sum or the mean.

This simple setup can model a wide variety of random phenomena, as we now demonstrate through four examples.

Example 1 (Roulette)**.** Suppose you want to know the probability of coming out ahead if you bet on "reds" on each of 100 spins of a roulette wheel. There are 38 spaces on a roulette wheel, all equally likely. If you bet on "reds", then 18 of the spaces (namely, the red ones) result in a gain of \$1, while the other 20 result in a loss of \$1. Therefore, we can model the net gain from a single bet on "reds" as a random draw from the box

$$\boxed{\underbrace{\boxed{+1}\cdots\boxed{+1}}_{\text{18 tickets}}\ \underbrace{\boxed{-1}\cdots\boxed{-1}}_{\text{20 tickets}}}. \tag{1}$$

Since we are betting on 100 spins of the roulette wheel, this is equivalent to drawing 100 times with replacement from the box (1). The net gain over those 100 spins is then the *sum* of the numbers drawn.

Example 2 (Simple Random Sample)**.** Suppose we take a simple random sample of size 400 from a town with 25,000 registered voters, of whom 12,000 are registered Republicans. What is the probability that more than half of the registered voters in our sample will be Republicans?

The population of 25,000 registered voters can be represented by the box

$$\boxed{\underbrace{\boxed{1}\cdots\boxed{1}}_{\text{12,000 tickets}}\ \underbrace{\boxed{0}\cdots\boxed{0}}_{\text{13,000 tickets}}}, \tag{2}$$

where a $\boxed{1}$ indicates a Republican and a $\boxed{0}$ some other political affiliation. The simple random sample can be represented as 400 draws from the box (2) without replacement. Since we are interested in the *proportion* of Republicans in the sample, we look at the *mean* of the numbers drawn. (The mean of a list of 0s and 1s is the proportion of 1s.)

From Example 2, it is a small leap to statistical inference. We assumed in that example that the number of Republicans in the town was known exactly, but if that were the case, there would be no point in collecting a sample in the first place. The point of collecting random samples is to *estimate* the number of Republicans in the population. For example, suppose in a random sample of 400 registered voters, we have 172 Republicans. What can we *infer* about how many Republicans are in the population based on this data?

We can rephrase the problem in terms of a box model. Now, we no longer know the exact composition of the box:

$$\boxed{\underbrace{\boxed{1}\cdots\boxed{1}}_{\text{??? tickets}}\ \underbrace{\boxed{0}\cdots\boxed{0}}_{\text{??? tickets}}},$$

but we do know that in 400 draws from this box, there were 172 $\boxed{1}$s. What can we *infer* about the tickets in the box based on these draws?

One way of making an inference is to conduct a hypothesis test, which assumes a composition for the box and determines if the observed data could plausibly have come from that box.

Example 3 (Hypothesis Test for a Simple Random Sample)**.** Suppose we take a simple random sample of size 400 from a town with 25,000 registered voters. We find that 172 of the people in our sample are Republicans. Based on this information, is it plausible that 45% of all registered voters in the town are Republicans?

Under the null hypothesis that 45% of the registered voters in the town are Republicans, we can write down the exact composition of the box:

$$\boxed{\underbrace{\boxed{1}\cdots\boxed{1}}_{\text{11250 tickets}}\ \underbrace{\boxed{0}\cdots\boxed{0}}_{\text{13750 tickets}}}.$$

Our sample is like the outcome of 400 random draws from this box without replacement. The only question is how likely are we to observe a result as extreme as 172 $\boxed{1}$s.

We can calculate the p-value, the probability of drawing 172 $\boxed{1}$s or fewer. This probability turns out to be about 0.225. Since this is quite large, the null hypothesis (that 45% of the registered voters in the town are Republican) appears to be consistent with observing 172 $\boxed{1}$s. On the other hand, if this probability had been very small, we would have had to conclude that the null hypothesis was untenable.

The next example illustrates how box models can reveal subtle distinctions in the data.

Example 4 (Toy Preference)**.** Hamlin, Wynn, and Bloom (2007) conducted a study in which 16 infants were shown two toy characters ("helper" and "hinderer") and asked to pick one. 14 infants picked the "helper" toy. Is this evidence that infants prefer the "helper" toy?

Under the null hypothesis of no preference, the infants should be equally likely to pick the "helper" and "hinderer" toys, so it is clear that the right test calculates how likely it is to observe 14 or more "helper" toys if the probability of choosing a "helper" toy is 0.5. But what exactly is the null hypothesis being tested? Rossman (2008) observes that there are actually two possible null hypotheses: (1) this data represent a random sample from a larger population of infants, of which half prefer the "helper" toy, or (2) each of the 16 infants in the study chooses the "helper" toy with probability 0.5, which we might observe if we were to test the same 16 infants repeatedly.

The box model clarifies the difference between the two scenarios. Under the first scenario, each infant in the population is predisposed to like either the "helper" (represented by $\boxed{1}$) or the "hinderer" (represented by $\boxed{0}$), and under the null hypothesis, there are a equal number of each in the population:

$$\boxed{\underbrace{\boxed{1}\cdots\boxed{1}}_{N/2 \text{ tickets}}\ \underbrace{\boxed{0}\cdots\boxed{0}}_{N/2 \text{ tickets}}}.$$

The data are like the outcome of 16 random draws from this box. The draws are made without replacement, but the distinction is irrelevant because the population size N is large relative to the sample size of 16. Under the second scenario, each infant randomly chooses one of the toys at random. If we represent the "helper toy" by a $\boxed{1}$ and the "hinderer" toy by a $\boxed{0}$, the data are like 16 random draws (with replacement) from the box

$$\boxed{\boxed{1}\ \boxed{0}}.$$

In either scenario, the number of $\boxed{1}$s can be assumed to follow a Binomial$(16, 0.5)$ distribution.

Although the mechanics of the hypothesis test are the same in both scenarios, the interpretations are very different. Under the first scenario, the results generalize to a larger population but require an assumption of random sampling. The second scenario requires no such assumption, but the results are specific to the 16 infants in the study. The box model exposes this subtle difference in interpretation because the boxes are manifestly different, even if the test is the same.

It is worth noting that the box model has appeared under various guises over the years. For example, Wood (2005) described random sampling using the "two bucket story", and Garfield, Zieffler *et al.* (2012) achieved a similar effect using the samplers and spinners in the software Tinkerplots$^{\text{TM}}$. Everything that we describe using box models can be reframed in terms of these alternative analogies, but we find the box model to be the most transparent way of describing and visualizing random sampling.

3. SIMULATION-BASED INFERENCE

So far, we have discussed only how to set up a box model, but not how to obtain probabilities and p-values from them. FPP ultimately obtain numerical answers from box models using normal approximations and formulas. For example, students are taught that

- the sum of n draws from a box approximately follows a normal distribution with

$$\text{EV} = n \times \text{average of the box} \qquad \text{SE} = \text{SD of the box} \times \sqrt{n} \tag{3}$$

- the average of n draws from a box approximately follows a normal distribution with

$$\text{EV} = \text{average of the box} \qquad \text{SE} = \frac{\text{SD of the box}}{\sqrt{n}}. \tag{4}$$

These normal approximations are sufficient to calculate any probability of interest, including p-values. One brilliant feature of these formulas is that they work for both means and proportions. In the traditional curriculum, students learn two SE formulas, one for means and another for proportions:

$$\text{SE}_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \qquad \text{SE}_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}. \tag{5}$$

The box model approach unifies the two formulas, by exploiting the fact that the (population) standard deviation of a collection of 0s and 1s is

$$\sigma = \sqrt{p(1-p)}. \tag{6}$$

However, students still struggle with these calculations because the formulas are unmotivated. For example, many students are confused about when to multiply by $\sqrt{n}$ in the SE formula and when to divide by $\sqrt{n}$. As Freedman *et al.* (2007) lamented,

> The students were willing to compute an SE as SD $\times \sqrt{n}$.... When they hit [the section on confidence intervals], there was a tremendous shifting of gears needed to compute the SE as SD$/\sqrt{n}$. Once they changed over, they stopped being able to compute the SE for a sum as SD $\times \sqrt{n}$—they insisted on dividing. We tried hard to explain that there was one formula to use with sums and another for averages, but they wouldn't buy it.

These calculations are only tangential to the logic of inference, yet they present a barrier for many students. Thus, FPP only partially achieved their goal of making sampling distributions transparent. Even though the process of setting up the box model is instructive, the calculations needed to obtain a final answer are not.

Cobb (2007) has argued persuasively that the technical difficulty of sampling distributions and normal approximations distracts students from the core conceptual ideas. As he eloquently puts it,

> There's a vital logical connection between randomized data production and inference, but it gets smothered by the heavy, sopping wet blanket of normal-based approximations to sampling distributions.

Cobb argues for a simulation-based approach to inference, which is accessible even to students who have not formally studied probability. As Moore (1997) noted earlier, "only an informal grasp of probability is needed to follow the reasoning of standard statistical inference." Many educators have taken up the banner, with some using simulation to introduce p-values as early as Week 1 of an introductory course (Roy, Rossman, Chance, Cobb, VanderStoep, Tintle, and Swanson 2014).

Our work provides a bridge between Freedman's box model analogy and simulation-based inference. In our experience, most courses that use the FPP textbook still rely exclusively on analytical calculations. To help instructors bridge the gap between box models and simulation-based inference, we developed a web application that enables simulation-based inference using the box model. Although there are other applications that can be used to carry out these sorts of simulations (e.g., the sampler in Tinkerplots^{TM}), ours is the only one that is explicitly tied to the analogy of tickets in boxes and is thus most natural for students learning from FPP. We have also adopted a number of ideas from user-centered design theory to help students scaffold their understanding of inference, as we describe in Section 4. Finally, we offer several extensions of the box model in Section 5, which to our knowledge, have not been pointed out in FPP or the literature.

4. THE BOX MODEL SIMULATOR

We have argued that the box model can be viewed as a computational engine that allows students to obtain inferences without detailed calculations. In this section, we demonstrate our implementation of this computational engine that we call the "box model simulator". The simulator is available online at

http://box-model.github.io/,

with source code available at

http://www.github.com/box-model.

The entire application was written in Javascript, so it can be run on any browser on any operating system, online or offline.

In designing this application, we adopted the design principle of *responsive disclosure* (Tidwell 2010), which says that detail should be hidden from users until they need to see it. We decomposed the box model analysis into four steps, with each step hidden from the user until the previous step is completed:

1. set up the box model and sample from the box (Figure 1)
2. aggregate the sample to produce a statistic (Figure 2)
3. repeat Steps 1-2 many times (Figure 3)
4. analyze the sampling distribution (Figure 4)

Figures 1-4 show the complete process for analyzing Example 4.

In the first step, we set up the box model by specifying (1) what tickets go in the box, (2) how many tickets to draw, and (3) whether the tickets should be drawn with or without replacement. (These numbers correspond to the speech bubbles in Figure 1.) There are three buttons on the side of the box for (4) adding tickets, (5) editing tickets, and (6) clearing the box. Clicking on either of the first two buttons reveal menus with additional options, which is yet another example of responsive disclosure. To continue to the next step, the user clicks on the "Sample" button (7), which reveals the "Sample" container and animates the sampling process.

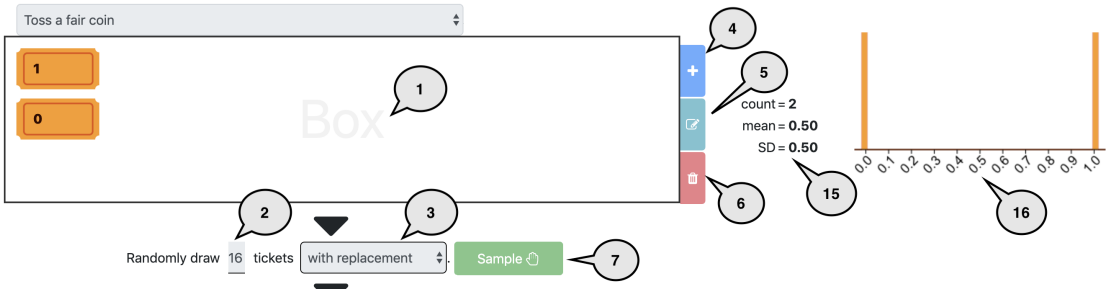

Figure 1: In the first step, we set up the box model and sample from it

Now that we have generated a sample, the next step is to aggregate the sample into a statistic. The application supports the sum and the mean (8), which are the same statistics that FPP uses. To complete this step, the user clicks the "Aggregate" button (9), which animates the aggregation process and reveals the "Statistic" container and a "Repeat" button. For the sample shown in Figure 2, the sum of the sixteen numbers is 10.

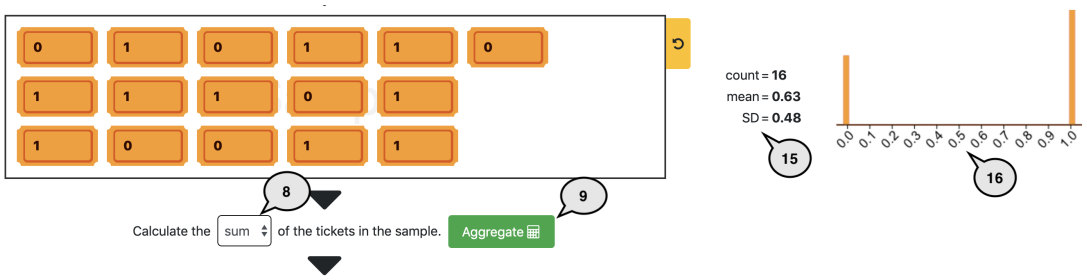

Figure 2: In the second step, we aggregate the sample numbers into a single statistic. In this example, the sum turns out to be 10. This statistic then appears in the box below (see Figure 3).

Clicking the "Repeat" button (10) reveals a menu that allows the user to repeat the first two steps many times. In this menu, the user is reminded explicitly what the procedure is and asked to specify how many times to repeat the procedure (11). Clicking the "Start" button (12) launches the animation, which makes it clear that the entire process is being repeated many times. In the example shown in Figure 3, we repeat the simulation 999 more times for a total of 1000 simulations.

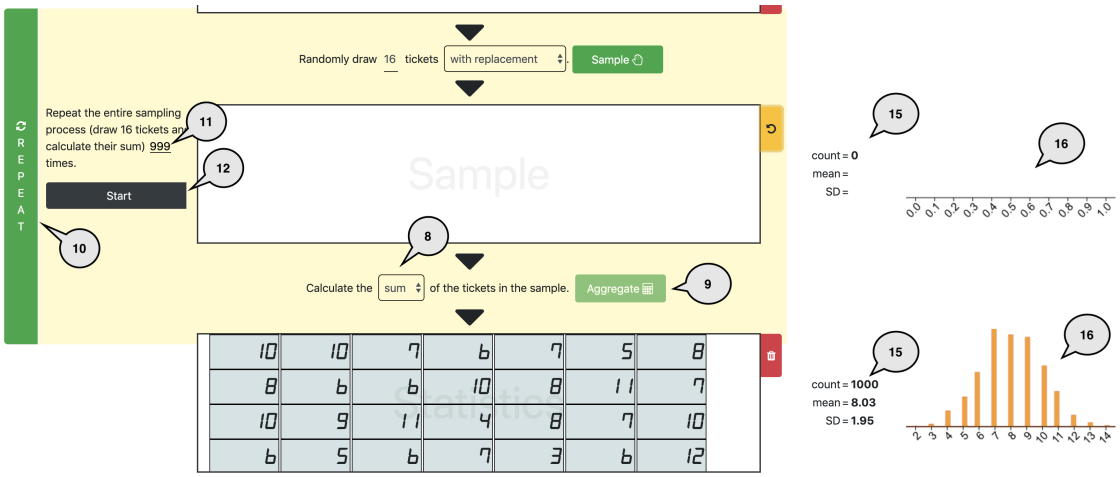

Figure 3: In the third step, we repeat the procedure many times to obtain a sampling distribution

Finally, now that we have a sampling distribution of 1000 statistics, we obtain p-values by simply counting the number of statistics that fell above or below (13) a particular threshold (14), usually the observed value of the statistic. In Example 4, 14 children chose the "helper" toy, so we count how often the statistic was greater than or equal to 14 in Figure 4. It only happened 2 times out of 1000, so the (approximate) p-value is 0.002 (or 0.2%).

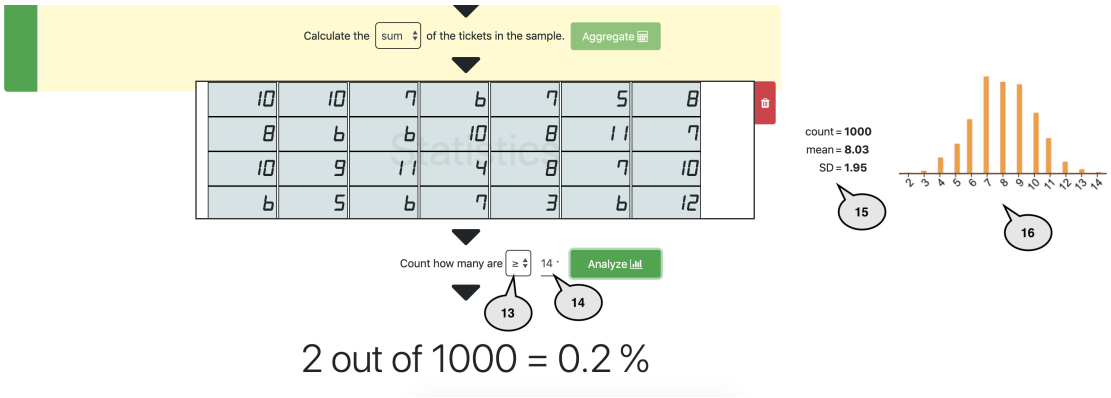

Figure 4: In the last step, we analyze the sampling distribution to obtain a p-value

The right side of the interface contains summary statistics (15)—the count, mean, and SD— and a histogram (16) of each collection of numbers. The summary statistics help students connect the results to the theoretical calculations. For example, students can verify empirically that the standard deviation of the means is indeed (SD of the box)/$\sqrt{n}$, as in equation (4). The histogram helps students visualize the Central Limit Theorem, since it will follow a normal distribution when the sample size is large.

Breaking down the process into these four distinct steps helps students conceptualize the logic of inference. For example, several authors (Saldanha and Thompson 2002; Case and Jacobbe 2018) have observed that students struggle to coordinate between the three levels of imagery: the population distribution, the distribution of the sample, and the sampling distribution. In our application, these three levels are represented by three distinct containers and histograms. At the same time, they are juxtaposed and aligned, to encourage comparison between them.

Like any online applet, the "box model simulator" is designed to enable a broad, but limited, set of simulations. For example, it only allows the user to use a predetermined set of statistics (e.g., sum, mean). Arbitrary statistics have to be specified using code. For a framework to specify and simulate from box models in Python, see Ross and Sun (2019).

5. USES OF THE BOX MODEL

In this section, we provide examples of lessons that can be taught using the box model.

5.1. Hypothesis Tests

We already alluded in Section 2 to the ways that FPP uses the box model. Tests of a point null hypothesis, such as of a binomial proportion, are easy to describe using the box model. For instance, the hypothesis test in Example 4 has an easy description in terms of the box

model.

While FPP would likely have carried out the hypothesis test using normal approximations, we advocate a three-pronged approach instead:

1. First, obtain the p-value by simulating from the box model using the applet described in Section 4.
2. Then, obtain the p-value using an exact binomial calculation. Section 5.3 describes how the binomial distribution can be introduced using the box model.
3. Finally, once the students have seen that the sampling distribution is approximately normal (through simulation and by inspecting a graph of the binomial p.m.f.), obtain the p-value using the normal approximation.

Unfortunately, other hypothesis tests are not so straightforward to carry out using a box model. For example, consider a one-sample test of a mean. The null hypothesis $\mu = \mu_0$ does not completely specify the contents of the box; in statistical parlance, it is a composite, rather than a point, null. Even if we adopt the standard assumption of the t-test that the tickets in the box follow a normal distribution with mean μ_0, the variance of this normal distribution remains an unknown nuisance parameter. In this situation, the box model remains a useful thought experiment, but it does not offer a prescription for simulating under the null. This shortcoming is not specific to the box model; other simulation-based curricula have also skirted the issue of how to simulate a one-sample test for a mean (Tintle, VanderStoep, Holmes, Quisenberry, and Swanson 2011; Diez, Barr, and Çetinkaya-Rundel 2014). Nevertheless, *approximate* inference using the box model is possible, provided that one is willing to use the bootstrap; see Section 5.2 for details.

FPP also try to situate two-sample tests in the framework of the box model. They distinguish carefully between the "population" and "randomization" models of two-sample inference (Lehmann 1975). In the "population" model, the data are regarded as two independent samples; FPP represents this situation by drawing independently from two boxes. In the "randomization" model, the randomness comes solely from the randomization of observations to treatment and control; FPP represents this situation using a single box, with double-sided tickets. We are sympathetic to the view of some educators that the box model analogy becomes strained at this point (Rossman and De Veaux 2016). Nevertheless, there are good simulation-based approaches to two-sample problems, such as permutation tests (Ernst 2004; Cobb 2007; Tintle *et al.* 2011), even though they do not fit neatly into the box model framework.

5.2. The Bootstrap

Most inference is ultimately based on the sampling distribution; that is, the distribution of the statistic under repeated sampling from the population. But we do not know the population; all we have is a single sample from the population. What can we learn about the sampling distribution from just one sample?

The bootstrap (Efron 1979) says that we can repeatedly resample from this sample to obtain "bootstrap samples". We can recalculate our statistic on these bootstrap samples to obtain a "bootstrap distribution" that allows us to learn about the sampling distribution. The bootstrap has a disarmingly simple description using the box model: we simply put the values from the sample data into a box and repeatedly sample from this box. See Figure 5 for a visualization of this process.

We can use the bootstrap to obtain one-sample inference for a mean. To evaluate the claim that $\mu = \mu_0$, we can construct a 95% confidence interval using the bootstrap and see if it contains μ_0. There are several ways to use the bootstrap samples to obtain a confidence interval; the simplest way is to simply take the 2.5% and 97.5% quantiles of the bootstrap distribution, but there are more complex methods with better coverage properties (Hesterberg 2015).

Alternatively, a bootstrap hypothesis test could be used to test a one-sample mean. This is the approach taken by Lock, Lock, Lock Morgan, Lock, and Lock (2013). For example, if we wanted to test the null hypothesis that $\mu = 3$ using the data in Figure 5, we could bootstrap using the sample data, after adjusting the data so that their mean is actually 3. Since the sample mean is $(7+2+4)/3 = 4.333$, we would have to subtract 1.333 from each of the sample observations to obtain a bootstrap population whose mean is $\mu = 3$. So in the end, we would draw bootstrap samples from the box

$$\boxed{\boxed{5.667}\ \boxed{0.667}\ \boxed{2.667}}.$$

As with all bootstrap procedures, the resulting inference is only an approximation, since the bootstrap population is not the same as the true population.

5.3. Defining the Discrete Distributions

Although the box model was designed for use in an introductory statistics class, it is potentially useful in more advanced classes as well. Here, we demonstrate another use of the box model in an introductory probability class, to define and unify the various discrete probability distributions.

Nearly all of the named discrete distributions can be defined in terms of a box model, using a box containing all 0s and 1s:

$$\boxed{\overbrace{\boxed{1}\cdots\boxed{1}}^{N_1 \text{ tickets}}\ \overbrace{\boxed{0}\cdots\boxed{0}}^{N_0 \text{ tickets}}}^{N \text{ tickets}}. \tag{7}$$

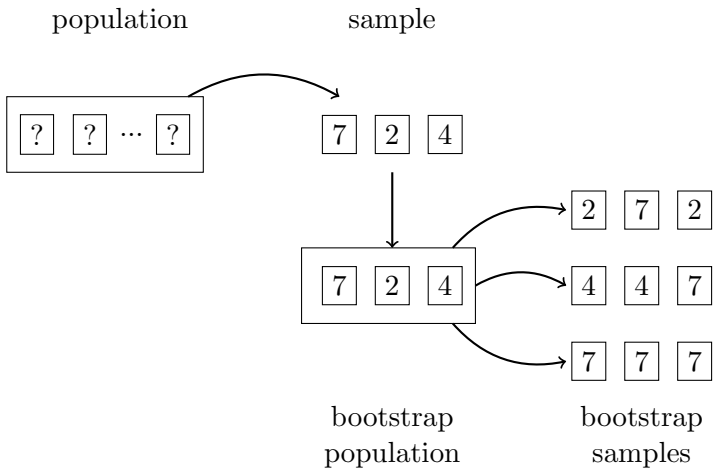

Figure 5: In most situations, we know little about the population from which our sample data (7, 2, and 4) were obtained. The bootstrap says we can simply put the sample values into a box and repeatedly resample from this box to obtain bootstrap samples.

Let $p \stackrel{def}{=} N_1/N$ be the proportion of $\boxed{1}$s in the box. Then,

- The outcome of 1 draw (i.e., either a 0 or a 1) is a Bernoulli(p) random variable.
- The number of $\boxed{1}$s in n draws *with* replacement is a Binomial(n, p) random variable.
- The number of $\boxed{1}$s in n draws *without* replacement is a Hypergeometric(n, N_1, N_0) random variable.
- The number of draws (with replacement) *until* we get a $\boxed{1}$ is a Geometric(p) distribution.
- The number of draws (with replacement) *until* we get r $\boxed{1}$s is a NegativeBinomial(r, p) random variable.

The only common discrete distribution that is conspicuously missing from the above list is the Poisson. But the Poisson can be defined as the approximate distribution when $N_1 \ll N$ and the number of draws n is large.

There are a number of advantages to introducing these distributions in this way.

- It emphasizes the connections between these distributions. It is not easy to appreciate just how closely the binomial and hypergeometric distributions are related when one is defined in terms of independent Bernoulli trials and the other is defined in terms of drawing black and white balls from an urn, as in most probability textbooks (Ross 2014; Blitzstein and Hwang 2014). Nor is their close connection obvious from their

p.m.f.s:

$$p[x] = \binom{n}{x} p^x (1-p)^{n-x} \qquad\qquad p[x] = \frac{\binom{N_1}{x}\binom{N_0}{n-x}}{\binom{N}{n}}.$$

But their descriptions in terms of the box model differ by just one word (*with* vs. *without* replacement)! Simulations from the box model can be used to illustrate that, for fixed n, N_1, and N_0, the binomial and hypergeometric distributions have the same center, but the hypergeometric distribution has less variability. The latter fact can be motivated intuitively: if you draw without replacement, each time you draw a $\boxed{1}$, you are less likely to draw a $\boxed{1}$ again.

- It immediately suggests new distributions that are outside the canon. For example, some students ask, "What if we drew *without* replacement in the definition of a negative binomial?" They have just independently discovered the "negative hypergeometric" distribution (Miller and Fridell 2007).

However, there is one downside to describing the discrete distributions using box models. For the distributions that involve drawing with replacement, a box model parametrized by N_1 and N_0 is unidentified. To see this, observe that the probability models obtained by drawing with replacement from the boxes

$$\boxed{\boxed{1}\ \boxed{0}} \qquad \text{and} \qquad \boxed{\boxed{1}\ \boxed{1}\ \boxed{0}\ \boxed{0}} \tag{8}$$

are identical, even though the boxes are different. When sampling with replacement, all that matters is the *relative proportion* of $\boxed{1}$s in the box, not the absolute number. This identifiability issue needs to be addressed in class, but it is unavoidable if one wishes to compare sampling with and without replacement from a population. In our experience, the insight from this comparison more than compensates for the trouble of an unidentifiable model.

Interestingly, FPP teaches the binomial distribution without using the box model, which is not introduced until later in the text. Thus, even though many of the hypotheses in the book could be tested exactly using binomial distributions, students miss this connection because the box model was not discussed in the context of the binomial distribution. In classes that cover both probability and statistics (such as engineering statistics courses), it may be fruitful to introduce the box model early so that students appreciate connections like using the binomial distribution to obtain exact p-values for testing proportions.

6. CONCLUSION

The box model analogy introduced by FPP can clarify difficult ideas in an introductory statistics course, such as sampling distributions and hypothesis testing. We have demonstrated how to enhance FPP's original presentation, by replacing normal approximations

with simulations. A simulation-based curriculum based on box models could be implemented using an interactive online applet we have developed that allows students to readily simulate from any box model and calculate probabilities. Finally, we have demonstrated additional uses (and limitations) of the box model beyond those pointed out by FPP.

We see our work as a technological bridge between two important innovations in statistics education: the box model analogy and simulation-based inference. We hope that our paper will inspire educators using the FPP textbook to incorporate simulation-based inference into their curriculum, as well as encourage educators using simulation-based inference to unify their course around the box model analogy.

ACKNOWLEDGEMENTS

Anelise Sabbag, Madeline Schroth-Glanz, two anonymous referees, and the editor offered suggestions that greatly improved this work. J.A. was supported by a Frost Undergraduate Student Research Award from the Bill and Linda Frost Fund.

REFERENCES

Blitzstein JK, Hwang J (2014). *Introduction to probability*. CRC Press.

Case C, Jacobbe T (2018). "A Framework to Characterize Student Difficulties in Learning Inference from a Simulation-Based Approach." *Statistics Education Research Journal*, **17**(2), 9–29.

Chance B, del Mas R, Garfield J (2004). "Reasoning about sampling distribitions." In *The challenge of developing statistical literacy, reasoning and thinking*, pp. 295–323. Springer.

Cobb GW (2007). "The introductory statistics course: A Ptolemaic curriculum?" *Technology Innovations in Statistics Education*, **1**(1).

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).

Diez DM, Barr CD, Çetinkaya-Rundel M (2014). *Introductory statistics with randomization and simulation*. OpenIntro.

Efron B (1979). "Bootstrap Methods: Another Look at the Jackknife." *The Annals of Statistics*, pp. 1–26.

Ernst MD (2004). "Permutation methods: a basis for exact inference." *Statistical Science*, **19**(4), 676–685.

Freedman D, Pisani R, Purves R (1978). *Statistics*. Norton.

Freedman D, Pisani R, Purves R (2007). *Instructor's manual for Statistics*. 4th edition. WW Norton.

Garfield J, Zieffler A, *et al.* (2012). "Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course." *ZDM*, **44**(7), 883–898.

Garfield JB, Ben-Zvi D, Chance B, Medina E, Roseth C, Zieffler A (2008). "Learning to Reason About Samples and Sampling Distributions." In *Developing Students' Statistical Reasoning*, pp. 235–259. Springer.

Hamlin JK, Wynn K, Bloom P (2007). "Social evaluation by preverbal infants." *Nature*, **450**(7169), 557.

Hesterberg TC (2015). "What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum." *The American Statistician*, **69**(4), 371–386.

Lehmann E (1975). *Nonparametrics: statistical methods based on ranks*. Holden-Day.

Lock PF, Lock RH, Lock Morgan K, Lock EF, Lock DF (2013). *Statistics: Unlocking the power of data*. Wiley.

Miller GK, Fridell SL (2007). "A forgotten discrete distribution? Reviving the negative hypergeometric model." *The American Statistician*, **61**(4), 347–350.

Moore DS (1997). "New pedagogy and new content: The case of statistics." *International statistical review*, **65**(2), 123–137.

Ross K, Sun DL (2019). "Symbulate: Simulation in the Language of Probability." *Journal of Statistics Education*, pp. 1–17.

Ross S (2014). *A first course in probability*. 9th edition. Pearson.

Rossman A (2008). "Reasoning about informal statistical inference: One statistician's view." *Statistics Education Research Journal*, **7**(2), 5–19.

Rossman A, De Veaux R (2016). "Interview With Richard De Veaux." *Journal of Statistics Education*, **24**(3), 157–168.

Roy S, Rossman A, Chance B, Cobb G, VanderStoep J, Tintle N, Swanson T (2014). "Using simulation/randomization to introduce p-value in week 1." In *Proceedings of the 9th International Conference on Teaching Statistics*, volume 9, pp. 1–6.

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

Tidwell J (2010). *Designing interfaces: Patterns for effective interaction design*. O'Reilly.

Tintle N, VanderStoep J, Holmes VL, Quisenberry B, Swanson T (2011). "Development and assessment of a preliminary randomization-based introductory statistics curriculum." *Journal of Statistics Education*, **19**(1).

Wood M (2005). "The role of simulation approaches in statistics." *Journal of Statistics Education*, **13**(3).