Teaching Materials and Classroom Practices for Simulation-Based Statistical Inference

Jimmy DOI, Mitsugu HASHIMOTO, Yasuhiko NAKAJIMA, Beth CHANCE,
Karen MCGAUGHEY, Soma ROY, Nathan TINTLE, Jill VANDERSTOEP, Michiko WATANABE

ABSTRACT

This paper provides an overview of two newly translated statistics lessons that focus on simulation-based inference (SBI). The topics are (1) inference for the difference of means (randomized experiment) and (2) inference for a mean difference (dependent samples). Lessons using SBI can be very effective to help students better understand statistical and mathematical concepts. Both lessons were recently presented at Super Science High Schools in Japan. We discuss the corresponding lesson materials, tactile simulations, computer simulations, and the lesson presentations. We also assess the lessons based on student survey feedback. Finally, we provide details on how to acquire the translated lesson materials via the teaching materials archive.

Keywords: Statistics education, simulation-based inference, active learning, curriculum, applets, hypothesis testing, independent samples, dependent samples, placebo effect

1. Introduction

In this paper we introduce two newly translated statistics lessons that use the simulation-based inference (SBI) approach for hypothesis testing. The first SBI lesson focuses on hypothesis testing for the difference in means for a randomized experiment, while the second SBI lesson addresses hypothesis testing for a mean difference with dependent samples. Both lessons, originally created by principal investigators of the STUB Network, were recently presented at two Super Science High Schools in Japan.

Section 2 provides background information related to the lessons. In Sections 3 and 4, we discuss the first and second SBI lessons, respectively. Section 5 is devoted to the teaching materials archive, and Section 6 includes concluding remarks.

2. Background Information

Before we delve into the specifics of the two lessons, we'd like to start with some relevant background information. First, we will touch on an important set of recommendations from the *Guidelines for Assessment and Instruction in Statistics Education* in the U.S. This will be followed by a brief introduction to simulation-based inference. Finally, we will shed light on the STUB Network.

(1) GAISE

Funded and endorsed by the American Statistical Association, the *Guidelines for Assessment and Instruction in Statistics Education* (GAISE) are frameworks for statistics education at both the college level (ASA, 2016) and the PreK-12 level (Bargagliotti et al., 2020). While the first GAISE college report was released in 2005, the updated college report released in 2016 contained the following six key recommendations:

1. Teach statistical thinking.
    a. Teach statistics as an investigative process of problem-solving and decision making.
    b. Give students experience with multivariable thinking.
2. Focus on conceptual understanding.
3. Integrate real data with a context and purpose.
4. Foster active learning.
5. Use technology to explore concepts and analyze data.
6. Use assessments to improve and evaluate

student learning.

Although this list appears in the college report, the recommendations are completely applicable to the high school setting as well. Since the release of the first GAISE college report, many statistics educators (at both the high school and college levels) across the United States and beyond have relied on these six key recommendations to guide their teaching practices.

GAISE continues to be an important topic in statistics education, and related publications cover topics such as preparing teachers to implement the GAISE recommendations (Garfield and Everson, 2009), incorporating GAISE recommendations into introductory statistics courses (Woodard and McGowan, 2012), and observing changes in students' attitudes towards statistics in GAISE-influenced introductory courses (Paul and Cunnington, 2017).

As we will show later, the lecture materials described in this paper are aligned with all six key GAISE recommendations, to varying extents.

(2) Introduction to SBI

Both lessons discussed in this paper are examples of using the SBI approach to teach statistical inference. As stated by Doi (2019), the SBI approach introduces students to the logic of statistical inference via simulations. Although the expression "simulation-based inference" emerged in statistics education publications after 2000, classroom practices involving use of simulation to illustrate key concepts were observed before 2000 (for example, see Scheaffer et al., 1996). Simulating sampling distributions and randomization distributions has been in practice in statistics education for at least 25 years, and the use of simulations beyond a single introductory lesson has been expanding (e.g., illustrating the logic of inference in comparing groups, regression) in curricula as well as statistical practice. One of the most compelling arguments made for centering curricula around simulation-based inference rather than the central limit theorem is given in Cobb (2007). In contrast to teaching statistical inference by first introducing formal probability and sampling distribution theory, an SBI approach begins with simple devices such as dice, coins, spinners, and cards to perform tactile simulations. In particular, research has shown that students gain a better understanding of abstract statistical concepts by first performing tactile simulations (Hancock and Rummerfield, 2020). Students then use the computer to increase the number of repetitions, enhancing the accuracy of empirical distributions, using those distributions to estimate p-values and confidence intervals.

According to Tintle et al. (2018), an important benefit of the SBI approach is that it is accessible to students with minimal statistical or mathematical background. In particular, Roy et al. (2014) state that SBI examples can be taught even during the first week of instruction of an introductory statistics course. These two points illustrate how SBI lessons can also be quite accessible at the high school level or even earlier (see Chance et al., 2023). Doi (2019) mentions that he successfully gave an SBI lesson (unrelated to the two lessons of this paper) to the students at Takasaki Super Science High School in Gunma Prefecture and none of the students had any prior statistics background.

When compared to students exposed to the traditional curriculum, studies have shown that there is improved conceptual understanding among students in courses that use SBI compared to those who use traditional curricula (Hildreth et al., 2018; Tintle et al., 2014; Chance et al., 2022). In addition, Mendoza and Roy (2018) found that students demonstrated discernably higher four-month retention of statistical concepts post-SBI courses compared to post-traditional courses; 16-month retention was found to be higher as well, though not statistically significantly.

As mentioned previously, a key component of SBI is the use of computer simulations, which can greatly enhance the understanding of concepts such as randomness, sampling, and variability (Chance et al., 2007; Garfield and Ben-Zvi, 2008). Doi (2019) describes applets from the Rossman/Chance Applet Collection (http://www.rossmanchance.com/applets/index2021.html) that are very useful for SBI-related lessons. The

simulation applets used for the lessons described in this paper are from this collection.

SBI continues to remain an active area of practice and research (see Burnham et al., 2023; Case et al., 2019). For resources on textbooks that use SBI, see Doi (2019). More recently, SBI-focused textbooks have been written by Tintle et al. (2019, 2020a, 2020b) and Lock et al. (2021). Finally, an SBI blog with many resources can be found at https://www.causeweb.org/sbi.

(3) STUB Network

As with almost all disciplines, the use of data and statistical thinking has now become a key part in the practice of biology. Biologists routinely rely on quantitative approaches for drawing conclusions from data. This has resulted in the inclusion of descriptive statistics and inferential statistical thinking in the typical undergraduate biology course (Brewer and Smith, 2011). The Statistical Thinking in Undergraduate Biology (STUB) Network was formed in 2017 to tackle the dearth of active discussion about teaching and assessment when integrating statistical thinking into biology courses.

Led by a team of experienced statistics and biology researchers and educators, the STUB Network is an online community of college-level biology and statistics instructors from a diverse set of institutions across the United States. Funded by the National Science Foundation (DUE-1730668), the STUB Network has facilitated conversations, learning opportunities, and sharing of ideas and materials.

These interdisciplinary conversations have resulted in a repository of free ready-for-in-class-use curricular materials. The network has developed over 50 peer-reviewed stand-alone activities/lessons for use in introductory and intermediate statistics courses and biology courses; all lessons are immersed in biology contexts, use real data, and utilize free-to-use web-based applets, thus improving both relevance and access. All curricular materials developed through this project remain freely accessible via the webpage: https://www.causeweb.org/stub.

The English version of the two lessons discussed in this paper are included in the STUB Network repository.

3. SBI Lesson 1: Inference for Difference of Two Means (Randomized Experiment)

(1) Introduction

The first SBI lesson is based on inference for the difference of two means in a randomized experiment The corresponding STUB Network materials can be accessed from https://bit.ly/STUB-INTRO (See Exploration 6.2). These materials were translated to Japanese by co-authors Doi, Hashimoto, and Nakajima.

The first presentation of this lesson was given at Hiroshima University Super Science High School and was team-taught by Doi and Hashimoto. The lesson spanned two 50-minute class sessions. The 34 participating students were in their second year.

Before going into the specifics of the lesson, we will first provide background information on the research study underlying the data analysis. Many animals in the wild can mentally track other nearby animals without seeing them. This mental tracking is a form of what is known as *socio-spatial cognition*. Takagi et al. (2021) performed a study to assess whether domesticated cats have socio-spatial cognition. In this experiment, each cat was placed in a room having two audio speakers that were at least four meters apart and were on opposite ends of the room. The cat's owner's voice was played through one of the speakers. After a pause of only a few seconds, either the (same) owner's voice or another person's voice was played from the other speaker. If domesticated cats have socio-spatial cognition, the researchers thought cats would exhibit a higher level of surprise when hearing their owner's voice from the second speaker as opposed to those hearing another person's voice. The reason for this is because, after hearing the owner's voice from the first speaker, cats may have a mental mapping of the owner being near that location. When they hear the owner's voice from the opposite end of the room through the second speaker, they might be quite surprised by the sudden spatial shift of the owner.

After hearing the voice from the second speaker, eight observers rated the cat's level of surprise on a scale from 0 (no surprise) to 4 (strongly surprised). The average of the eight ratings was recorded as the cat's "surprise score." A total of 40 cats participated in the study and the voice condition was randomly assigned to each cat.

(2) Experiment Data and Hypotheses

Of the 40 cats, 21 were randomly assigned to hear their owner's voice ("same" group) from the second speaker and the remaining 19 were assigned to hear another person's voice ("diff" group) from the second speaker. Based on the experiment, the difference in surprise score means ($\bar{x}_{\text{same}} - \bar{x}_{\text{diff}}$) was 0.394.

Let μ_{same} be the population mean surprise score for cats that hear the same voice (the owner's) from the second speaker and let μ_{diff} denote the population mean surprise score for cats that hear a different voice from the second speaker. The null and alternative hypotheses to investigate whether hearing the same voice has an effect on surprise level of the cats can be written as follows:

H_0: $\mu_{\text{same}} = \mu_{\text{diff}}$ H_a: $\mu_{\text{same}} \neq \mu_{\text{diff}}$

(Note: Although the researchers had a one-sided alternative hypothesis in mind, we used the more conservative two-sided test in this lesson.)

If the null hypothesis is true, then the voice type from the second speaker should have no effect on surprise score, on average. In particular, if the null hypothesis is true, then reactions by the cats would have been *unchanged* by whether the owner's voice or another voice was played. This will be an important consideration for the in-class simulation.

(3) Class Activity: Card Shuffling

Based on the difference in means observed in the experiment (0.394), the key question we posed to the students was "How often would such an extreme difference occur by chance alone if the null hypothesis were true?" We addressed this by performing a tactile simulation in class.

Assuming that the null hypothesis is true (voice condition has no impact on surprise score), we can perform a corresponding simulation by randomly re-assigning the 40 observed surprise scores between the two voice groups.

For the simulation, we distributed to each student a set of 40 cards. The cards contained the surprise scores observed in the experiment. Figure 1 shows an image of the cards template. The corresponding PDF file (available from the teaching materials archive, see Section 5) can be used to print the cards for this activity.

We asked the students to shuffle the cards and to designate 21 for the "same" group and 19 for the "diff" group to match the study. An example of the random assignment using the cards is shown in Figure 2.

0.375	0.25	0.75	1	3.5	0.625	3.25	0.625
1.875	4	1.5	1.375	2	1	2.25	1
2.375	2.125	0.875	0.375	0.5	2.5	0.375	0.625
0.75	0.875	1.375	2.25	2.125	2.75	2.875	2
1.375	1.375	2.25	1.375	2.25	2.75	2.75	1.125

Figure 1. Image of cards template containing all 40 surprise scores.

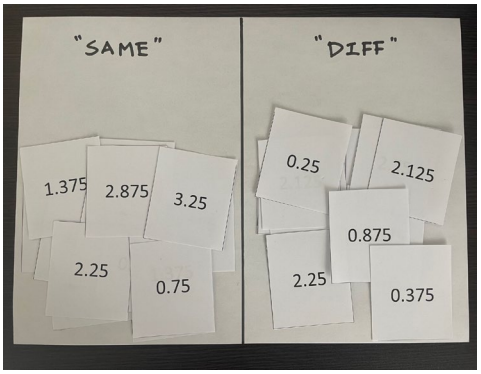

Figure 2. Example where cards are assigned randomly to the two groups.

We then had students determine the means for their re-randomized "same" and "diff" groups ($\bar{x}_{\text{same}}$ and $\bar{x}_{\text{diff}}$, respectively) and to compute the simulated difference in means ($\bar{x}_{\text{same}} - \bar{x}_{\text{diff}}$). The differences in means from

the students' tactile simulation were combined to construct a histogram of the simulated null distribution of difference in means, with the experimental outcome of 0.394 overlaid on the histogram, as shown in Figure 3.

The students seemed to understand that this histogram reveals what we could expect to see for the difference in means if the null hypothesis were true. Next, we posed the key question of whether the experimental outcome of 0.394 seemed quite extreme relative to this histogram. Most students agreed that the outcome 0.394 did not appear to be particularly extreme. They also agreed that more repetitions would be helpful to assess the question at hand.

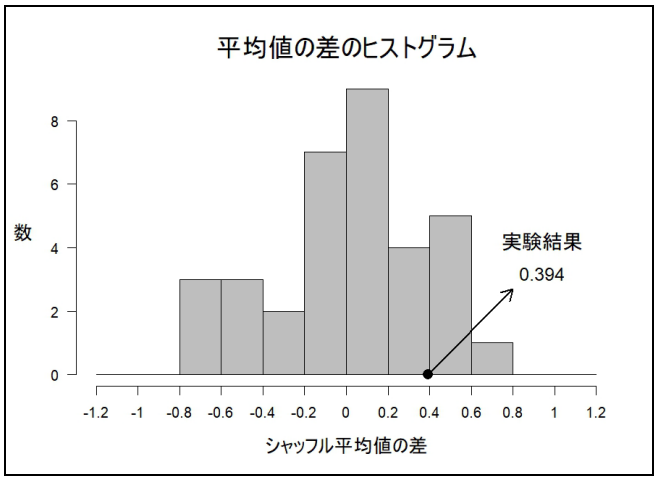

Figure 3. Histogram of the difference in means from 34 students' tactile simulations. The experimental outcome (0.394) is overlaid.

(4) Simulation Applet

To increase the number of repetitions, we used an applet from the Rossman/Chance Applet Collection. Note that the applets from this collection can be used on laptops and on other mobile devices such as tablets and smartphones.

The specific applet we used, which was translated to Japanese, can be accessed from https://bit.ly/applet-means. The applet simulates the card shuffling activity and generates a corresponding histogram of the difference of means.

Initially, each student used the applet to generate a histogram for 34 repetitions (matching the number of student repetitions from the tactile simulation). Students noted that the histograms from their simulations greatly varied from one another. As the students increased the number of repetitions (to 400, 4,000, 8,000), they noted their corresponding histograms began to converge. After generating 10,000 repetitions, the students used the applet to determine the proportion of outcomes that were as extreme or more extreme (in either direction) than the experimental outcome of 0.394. This proportion can be thought of as an approximate two-sided p-value for the randomization test.

Figure 4 shows a screenshot of the applet based on 10,000 repetitions, with 0.1949 as the corresponding proportion of outcomes as or more extreme than 0.394. Although this proportion will vary from student to student based on their simulations, all should be similar to 0.19. Due to the size of this proportion, most students recognized that the experimental outcome of 0.394 is not particularly extreme relative to the histogram of the null distribution. Based on this, many students realized there is not strong evidence against the null hypothesis and therefore the null hypothesis should not be rejected[1].

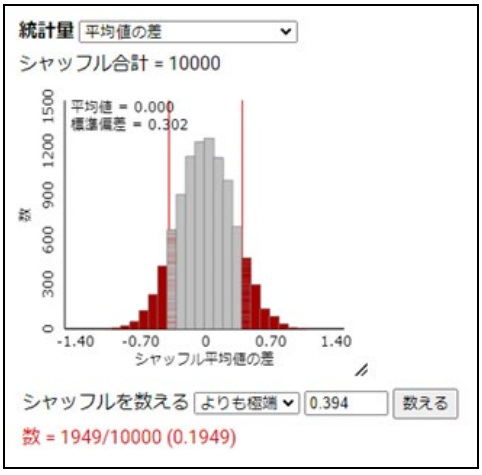

Figure 4. Simulation applet screenshot. Display shows the distribution of the difference in means from 10,000 repetitions and the proportion of outcomes at least as extreme as the experiment outcome of 0.394.

Based on the computer simulation's ease of use, it may be tempting to have students skip the tactile

[1] Takagi et al. analyzed a larger data set and based on a different testing procedure their results supported the claim that cats have socio-spatial cognition.

simulation and only use the applet, but we advise against this approach. As with Chance et al. (2007) and Justice et al. (2020), we believe there is a greater benefit for students when they engage in tactile simulations prior to utilizing technology. A related study by Hancock and Rummerfield (2020) demonstrated that students who participated in tactile simulations before doing computer simulations performed better on a subsequent assessment when compared to those who exclusively used computer simulations. Regarding tactile simulations, Hancock and Rummerfield stated, "Though these hands-on activities may take valuable class time, their use appears to be beneficial for the student, and can be rewarding and enjoyable for the instructor" (Hancock and Rummerfield, 2020, p. 16).

(5) Survey Results

At the conclusion of this lesson, students were asked to complete an anonymous survey. Each survey question was based on a 4-point Likert scale where 1 = very positive, 2 = positive, 3 = negative, and 4 = very negative. We dichotomized the responses as "positively disposed" (1 or 2) and "negatively disposed" (3 or 4). The following response percentages were based on this dichotomization. A summary of the survey results (based on 34 participants) is shown below:

- 91% found the corresponding research study to be interesting.
- 85% felt they understood how the experiment was conducted and how the explanatory and response variables tied-in to the corresponding research question.
- 76% felt they understood why the random assignment of observed surprise scores (via card shuffling) was consistent with the null hypothesis being true.
- 88% felt the applet was easy to use.
- 82% felt they understood why the simulation histograms were centered at zero and bell-shaped.
- 88% found the lesson to be worthwhile.

The students also had an opportunity to provide anonymous free response feedback. Here are some of their responses:

- "Going through the process of hypothesis testing from the beginning based on an actual study was great as it helped me understand why each step was performed."
- "It was good that we were able to verify that the histograms were bell-shaped rather than using something like a normal distribution out of the blue."
- "I want to make use of what we learned in an actual research project."
- "There were so many technical terms I found it difficult to follow at times."

Based on the survey results, overall, we felt that the first presentation of SBI Lesson 1 was quite successful. Although the vast majority of the free response feedback was positive, we noted that a couple students left comments indicating they found the lesson difficult to follow. We believe this may be due to the pacing of the lesson and we reflect on this point next.

(6) Lesson Reflections

Although we had two 50-minute class sessions for the lesson, we felt that was not enough time in class to go through all 19 questions from the original lesson plan. So, we skipped some of the more tangential questions from the notes related to confidence intervals and inference based on the median.

Even after omitting some questions, it was a challenge to cover the rest of the material in the allotted time. Perhaps one of the greatest challenges was the explanation of the cat experiment to the students. Finding an appropriate balance of study background to present that will engage, but not overwhelm, students will depend on the students' backgrounds. To address this point, next time we think we could have students read about the experiment details outside of class prior to the lesson. Another option would be to refine the explanation and omit details that are not directly related to the hypothesis test, such as socio-spatial cognition.

A few months after the initial presentation, Hashimoto presented SBI Lesson 1 again (on his own) at Hiroshima University Super Science High School. Around the same time, Nakajima presented SBI Lesson 1 (on his own) at Maebashi Super Science High School. With insights learned based on the initial presentation, the subsequent lessons seemed to go more smoothly.

4. SBI Lesson 2: Inference for Mean Difference (Dependent Samples)

(1) Introduction

The second SBI lesson is based on inference for the mean difference with dependent samples. The corresponding STUB Network materials can be accessed from https://bit.ly/STUB-INTRO (See Exploration 7.1-7.2). These materials were translated to Japanese by co-authors Doi, Hashimoto, and Nakajima.

The first presentation of this lesson was given at Maebashi Super Science High School and was team-taught by Doi and Nakajima. The lesson spanned two 50-minute class sessions. The 26 participating students were in their second year.

Before going into the specifics of the lesson, we will again provide background information on the research study underlying the data analysis. Simply rinsing your mouth with a carbohydrate solution while running has been shown to enhance performance. Researchers Brown et al. (2021) aimed to determine whether these rinses have a placebo effect. In particular, they investigated whether this performance enhancement would be increased if the solution was dyed pink compared to a clear solution, as previous research suggests that the color pink is associated with greater perceived sweetness.

A total of 10 runners participated in the study. Each participant ran on a treadmill for 30 minutes. Two non-caloric artificially sweetened solutions were prepared for each runner. One was dyed pink, and the other was left clear. The runners rinsed their mouths out with the solution, randomly assigned to be pink or clear. The runners repeated this rinse every 5 minutes during their runs. One week later, all the participants returned to repeat the experiment but rinsing with a solution of the other color (clear or pink) than the one they were originally assigned. The distance each participant ran (measured in meters) for each 30 minute-session was recorded.

Given that each runner performed the experiment twice, their distance based on the pink solution would be associated with their distance based on the clear solution. Due to this association, we say that the two samples (pink distance, clear distance) are dependent. Given that the response variable values arise in pairs, this type of study is often referred to as a *paired design*.

(2) Experiment Data and Hypotheses

The distances from the experiment are shown in Table 1. Because the data are paired, we compare the two distances for each runner by calculating the difference in distances between the two solutions. The last column of Table 1 contains the difference in distances (pink – clear). The mean difference from the experiment was 211.9m.

Table 1. Distances (in meters) for each of the 10 participants. Also shown are the difference in distances and the mean difference (211.9m).

参加者	ピンク色 溶液距離	透明 清液距離	ピンク色－透明
1	4105	3483	622
2	4361	3862	499
3	4105	4172	－67
4	4828	4758	70
5	4845	4791	54
6	4845	4995	－150
7	5205	5062	143
8	5912	5443	469
9	5827	5702	125
10	6440	6086	354
			平均差 ＝211.9

Let μ_d be the population mean difference in running distance when rinsing with a pink solution and clear solution (pink – clear). The null and alternative hypotheses to investigate whether solution color has an effect on running distance are as follows:

H_0: $\mu_d = 0$ H_a: $\mu_d \neq 0$

(Note: Although the researchers had a one-sided

alternative hypothesis in mind, we used the more conservative two-sided test in this lesson.)

If the null hypothesis is true, then solution color should have no effect on running distance, on average. If this is true, then distances for each runner would have been *unchanged* if the solution colors were switched. This will be an important consideration for the in-class simulation.

(3) Class Activity: Coin Flipping

Based on the mean difference from the experiment (211.9m), we again posed the key question "How often would such an extreme difference occur by chance alone if the null hypothesis were true?" As done in the previous lesson, we addressed this question by performing a tactile simulation in class.

Assuming that the null hypothesis is true (solution color has no impact on distance), we can perform a corresponding simulation by randomly assigning solution color for the pair of distances for each runner. This can be done with a coin toss. If the coin lands heads, then the solution color designation is switched compared to the actual experiment. Otherwise, the solution color designation remains unchanged. An example of the simulation is shown in Table 2.

Table 2. Simulation example based on 10 random tosses of a coin. If the coin lands "heads" then distances are switched. Otherwise, the distances are not switched. The mean difference for this simulation example is 67.7m.

コイン投げ（表＝交換）	ピンク色溶液距離	透明清液距離	ピンク色－透明
表	3483	4105	－622
裏	4361	3862	499
裏	4105	4172	－67
表	4758	4828	－70
表	4791	4845	－54
表	4995	4845	150
裏	5205	5062	143
裏	5912	5443	469
表	5702	5827	－125
裏	6440	6086	354
			平均差 ＝67.7

We asked the students to follow the example of Table 2 and perform their own simulation by tossing a coin. We then asked the students to determine the difference in distances (pink – clear) and to compute the mean difference ($\bar{x}_d$). Based on their simulations, we constructed a histogram of the simulated null distribution of mean difference and we also overlaid the experimental outcome of 211.9. This histogram is shown in Figure 5.

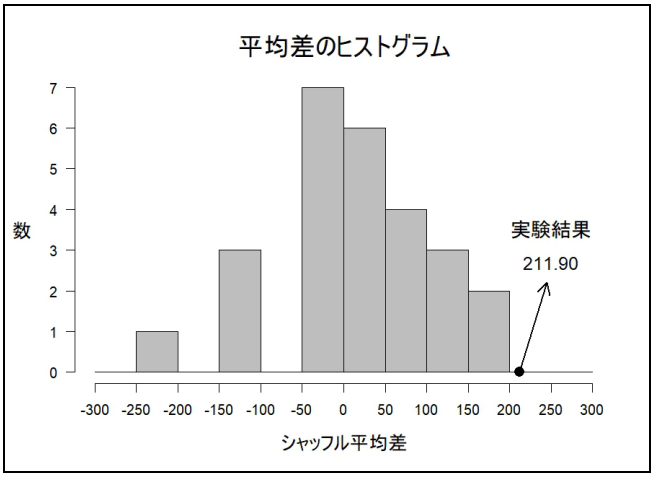

Figure 5. Histogram of the mean difference from 26 students' tactile simulations. The experimental outcome (211.90) is overlaid.

The students were reminded that this histogram reveals what we could expect for the mean difference if the null hypothesis were true. We repeated the key question of whether the experimental outcome of 211.9 seemed extreme relative to this histogram. Noting that no simulation results exceeded 200, all students agreed that the experimental outcome appeared to be extreme. Due to the low number of repetitions performed, the students suggested that more repetitions would be helpful before drawing any conclusions.

(4) Simulation Applet

To increase the number of repetitions, we again used an applet from the Rossman/Chance Applet Collection. The specific applet we used, which was translated to Japanese, can be accessed from https://bit.ly/applet-pairs. The applet simulates the coin tossing activity and generates a corresponding histogram of the mean difference.

Initially, each student used the applet to generate 26 repetitions (matching the number of student repetitions

from class). Comparing their simulation outcomes, students noted that the histograms they generated varied greatly from one another. As the students increased the number of repetitions (to 400, 4,000, 8,000) they noted their corresponding histograms began to converge. After generating 10,000 repetitions, the students used the applet to determine the proportion of outcomes that were as extreme or more extreme (in either direction) than the experimental outcome of 211.9. Again, this proportion can be thought of as an approximate two-sided p-value for the test.

Figure 6 shows a screenshot of the applet based on 10,000 repetitions, with 0.0326 as the corresponding proportion of outcomes as or more extreme than 211.9. Although this proportion will vary from student to student based on their simulations, all should be similar to 0.03. Due to the small size of this proportion, most students recognized that the experimental outcome of 211.9 is extreme relative to the histogram of the null distribution. Based on this, many students felt there is sufficient evidence against the null hypothesis and therefore the null hypothesis should be rejected.

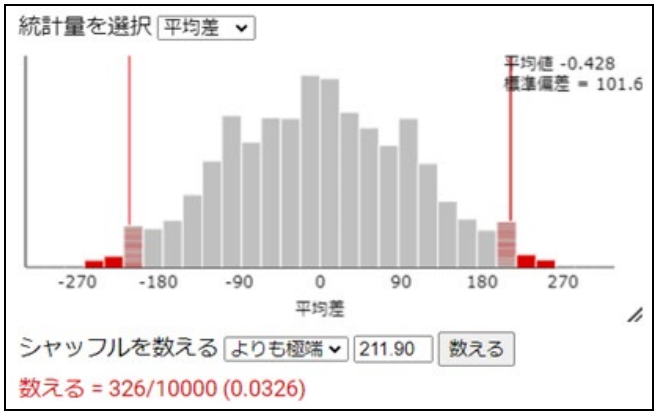

Figure 6. Simulation applet screenshot. Display shows the distribution of the mean difference from 10,000 repetitions and the proportion of outcomes at least as extreme as the experiment outcome of 211.9.

Here, we would like to mention once again (as was stated at the end of Section 3.4) that, although exclusively using the computer simulation may save class time, we recommend that students perform the tactile simulation as well to help them understand how the simulation process changes to match the study design.

(5) Survey Results

We again asked students to complete an anonymous survey concerning the lesson. The response percentages were based on the same 4-point Likert scale and dichotomization that were used in the previous survey. A summary of the survey results (based on 26 participants) is shown below:

- 96% thought the corresponding research study to be interesting.
- 100% felt they understood how the experiment was conducted and how the explanatory and response variables tied-in to the corresponding research question.
- 100% felt they understood why the random assignment of the observed running distances (via coin tossing) was consistent with the null hypothesis being true.
- 100% felt the applet was easy to use.
- 96% felt they understood why the simulation histograms were centered at zero and bell-shaped.
- 100% found the lesson to be worthwhile.

The students also had an opportunity to provide anonymous free response feedback. Here are some of their responses:

- “The simulation applet was very easy to use, and concepts that were difficult to visualize were easier to understand by creating my own diagrams.”
- “I wish the contents of our math textbooks would be more like this lesson.”
- “I found the lesson to be meaningful as I was able to learn testing methods that can be directly used in my own research.”
- “There were times when I thought it was difficult to understand because there was a lot of information.”

Based on the survey results, overall, we felt that the first presentation of SBI Lesson 2 was quite successful. Again, most free response feedback was positive. We noticed that, as with the previous survey, a couple students expressed concern about the amount of

information presented. We again believe this was due to the pacing of the lesson and we reflect on this point next.

(6) Lesson Reflections

As was the case with SBI Lesson 1, we felt we did not have enough class time to go through all questions from the original lesson notes (32 total) for the two 50-minute class sessions. Part 1 of the lesson notes introduces the concept of the paired design. In place of Part 1 we prepared a more succinct discussion to introduce this concept. Also, we skipped some of the more tangential questions from the notes related to confidence intervals and inference assuming the samples were independent.

In retrospect, we may have hurried through the explanation of paired designs and the experiment. This may be the underlying reason why a couple students found the lesson difficult to follow. To address this point we believe for the next iteration we could have students read about paired designs and/or experiment details outside of class prior to the lesson.

A few months after the initial presentation, Nakajima presented SBI Lesson 2 again (on his own) at Maebashi Super Science High School. Around the same time, Hashimoto presented SBI Lesson 2 (on his own) at Hiroshima University Super Science High School. After making some adjustments based on the initial presentation, the subsequent lessons seemed to be go more smoothly.

5. Teaching Materials Archive

(1) SBI Lesson 1: Inference for Difference of Two Means (Randomized Experiment)

All teaching materials related to the first SBI lesson have been archived in a ZIP file. The archive file can be accessed from https://bit.ly/SBI-ARCHIVE or from https://bit.ly/STUB-INTRO (See Exploration 6.2). The contents of the archive file are described below:

a) Lesson Notes: (9 pages, in Japanese)
b) Solution File: (in Japanese) Solutions to questions from lesson notes (access granted only to verified educators)
c) Lesson Plan: (in Japanese) Detailed guide of the objectives, methods, materials, and time schedule for the lesson.
d) Image Files for Lesson: Folder of images that can be used during the lesson.
e) Template File for Cards: File can be used to print cards for the corresponding tactile simulation. Contains all 40 surprise scores from the original cat experiment.
f) Excel Simulation File: (in Japanese) This Excel program performs a single repetition mirroring the card shuffling activity. Although we recommend students perform the tactile simulation themselves, if there is insufficient time for students to do this, the Excel program can be used instead.
g) URL for Simulation Applet (Japanese translated)

(2) SBI Lesson 2: Inference for Mean Difference (Dependent Samples)

All teaching materials related to the second SBI lesson have also been archived in a ZIP file. The archive file can be accessed from https://bit.ly/SBI-ARCHIVE or from https://bit.ly/STUB-INTRO (See Exploration 7.1-7.2). The contents of the archive file are described below:

a) Lesson Notes (13 pages, in Japanese)
b) Solution File: (in Japanese) Solutions to questions from lesson notes (access granted only to verified educators)
c) Lesson Plan: (in Japanese) Detailed guide of the objectives, methods, materials, and time schedule for the lesson.
d) Image Files for Lesson: Folder of images that can be used during the lesson.
e) Excel Computation File: This Excel program can be used to calculate the average of the differences of the shuffled distance values generated from the tactile simulation.
f) Excel Simulation File: (in Japanese) This Excel

program performs a single repetition mirroring the coin tossing activity. Although we recommend students perform the tactile simulation themselves, if there is insufficient time for students to do this, the Excel program can be used instead.

g) URL for Simulation Applet (Japanese translated)

6. Conclusion

At this point, SBI Lesson 1 and 2 have been presented multiple times at Hiroshima Super Science High School and Maebashi Super Science High School, with good success. Using a simulation-focused approach, these lessons offer an effective way to present the reasoning of statistical inference in a very accessible manner. Although these lessons have been presented at high schools, the materials are also perfectly suitable for introductory statistics courses at universities as well.

Another benefit of these lessons is that they both satisfy all six key GAISE recommendations (mentioned in Section 2.1), to varying extents. We provide the recommendations again along with an explanation of how they tie-in to the lessons:

1. Teach statistical thinking – Both lessons focus on the investigative process of problem-solving and decision making.
2. Focus on conceptual understanding – Instead of emphasizing formulas and computations, both lessons concentrate more on understanding concepts such as randomness, variability, distributions, and the hypothesis testing process.
3. Integrate real data with a context and purpose – Both lessons analyze real data sets and published studies and provide full context of the underlying research problem.
4. Foster active learning – Both lessons utilize tactile and computer simulations that require students to actively participate.
5. Use technology to explore concepts and analyze data – Both lessons rely on simulation applets to visualize the randomization process and collect results that are used to approximate p-values.
6. Use assessments to improve and evaluate student learning – Both lessons have corresponding notes filled with questions that are designed to help students understand statistical concepts.

Comparing the SBI approach to a traditional curriculum, Doi (2019) stated “It is important to point out that SBI methods need not replace traditional teaching content and methods but can instead *enhance* them” (p. 31). Both Hashimoto and Nakajima found this to be true for their respective classes. Their presentations of SBI Lesson 1 and 2 preceded presentations that introduced formal inference using a traditional curriculum. Thanks to the foundation of the two SBI lessons, both Hashimoto and Nakajima felt that their students were better able to understand concepts such as sampling distribution, p-value, significance level, and the overall hypothesis testing process. The faculty at both Hiroshima Super Science High School and Maebashi Super Science High School are planning to adopt SBI Lesson 1 and 2 as part of their math curricula.

REFERENCES

American Statistical Association. (2016). *Guidelines for Assessment and Instruction in Statistics Education (GAISE): College report*. Alexandria, VA: American Statistical Association. https://www.amstat.org/asa/files/pdfs/GAISE/GaiseCollege_Full.pdf.

Bargagliotti, A., Franklin, C., Arnold, P., Gould, R., Johnson, S., Perez, L., & Spangler, D. (2020). *Pre-K-12 Guidelines for Assessment and Instruction in Statistics Education (GAISE) report II*. American Statistical Association and National Council of Teachers of Mathematics.

Brewer, C., and Smith, D. (2011). *Vision and Change in Undergraduate Biology Education: A Call to Action*. https://www.aps.org/programs/education/undergrad/upload/Revised-Vision-and-Change-Final-Report.pdf. PDF file.

Brown, D., Cappozzo, F., De Roeck, D., Zariwala, M.. and Deb, S. (2021). “Mouth Rinsing with a Pink Non-caloric, Artificially-Sweetened Solution

Improves Self-Paced Running Performance and Feelings of Pleasure in Habitually Active Individuals," Front. Nutr. 8:678105. doi: 10.3389/fnut.2021.678105

Burnham, E., Blankenship, E., and Brown, S. (2023). "Designing a Large, Online Simulation-Based Introductory Statistics Course," Journal of Statistics and Data Science Education, 31:1, 66-73, DOI: 10.1080/26939169.2022.2087810

Case, C., Battles, M., and Jacobbe, T. (2019). "Toward an understanding of p-values: Simulation-based inference in a traditional statistics course," Investigations in Mathematics Learning, 11:3, 195-206, DOI: 10.1080/19477503.2018.1438869

Chance, B., Ben-Zvi, D., Garfield, J., and Medina, E. (2007). "The Role of Technology in Improving Student Learning of Statistics," Technology Innovations in Statistics Education, 1(1).

Chance, B., Medina, E., and Silverbush, J. (2023). *If You Only Have One Hour … Teaching Statistical Inference to Youth*. Statistics Teacher 23 Mar. 2023, https://www.statisticsteacher.org/2023/03/23/teaching-statistical-inference/. Accessed 15 APR 2024.

Chance, B., Tintle, N., Reynolds, S., Patel, A., Chan, K., Leader, S. (2022). "Student Performance in Curricula Centered on Simulation-Based Inference," Statistics Education Research Journal, 21(3), Article 4. https://doi.org/10.52041/serj.v21i3.6 Cobb, G. W. (2007). The Introductory Statistics Course: A Ptolemaic Curriculum? Technology Innovations in Statistics Education, 1(1). http://dx.doi.org/10.5070/T511000028

Cobb, G. (2007). "The Introductory Statistics Course: A Ptolemaic Curriculum?," Technology Innovations in Statistics Education, 1(1). http://dx.doi.org/10.5070/T511000028 Retrieved from https://escholarship.org/uc/item/6hb3k0nz

Doi, J.（2019）シミュレーションに基づく推論とアクティブ・ラーニングの授業事例，日本数学教育学会誌，101，3，28-39.

Garfield, J., and Ben-Zvi, D. (2008). *Developing Students' Statistical Reasoning: Connecting Research and Teaching Practice*, Kluwer Academic Publishers.

Garfield, J., and Everson, M. (2009). "Preparing Teachers of Statistics: A Graduate Course for Future Teachers," Journal of Statistics Education, 17:2, DOI: 10.1080/10691898.2009.11889516

Hancock, S., and Rummerfield, W. (2020), "Simulation methods for teaching sampling distributions: Should hands-on activities precede the computer?," Journal of Statistics Education. 28:1, 9-17, DOI: 10.1080/10691898.2020.1720551

Hildreth L., Robison-Cox J., and Schmidt J. (2018). "Comparing Student Success and Understanding in Introductory Statistics under Consensus and Simulation-Based Curricula," Statistics Education Research Journal, 17(1), 103-120. https://iase-web.org/documents/SERJ/SERJ17(1)_Hildreth.pdf.

Justice, N., Le, L., Sabbag, A., Fry, E., Ziegler, L., and Garfield, J. (2020). "The CATALST Curriculum: A Story of Change," Journal of Statistics Education, 28(2), 175–186. https://doi.org/10.1080/10691898.2020.1787115

Lock, R, Lock, P, Morgan, K, Lock, E, and Lock, D. (2021). *Statistics: Unlocking the Power of Data* (3rd Ed), Wiley.

Mendoza, S., and Roy, S. (2018). "Assessing Retention of Statistical Concepts after Completing a Post-Secondary Introductory Statistics Course," Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan https://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_C177.pdf

Paul, W., and Cunnington, R. C. (2017). "An Exploration of Student Attitudes and Satisfaction in a GAISE-Influenced Introductory Statistics Course," Statistics Education Research Journal, 16(2), 487-510

Roy, S., Rossman, A., Chance, B., Cobb, G., VanderStoep, J., Tintle, N., and Swanson, T. (2014). "Using Simulation/Randomization to Introduce P-Value in Week 1," Proceedings of the Ninth International

Conference on Teaching Statistics. http://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_4A2_ROY.pdf.

Scheaffer, R., Gnanadesikan, M., Watkins, A., and Witmer, J. A. (1996). *Activity-Based Statistics*, New York: Springer Verlag.

Takagi, S., Chijiiwa, H., Arahori, M., Saito, A., Fujita, K., and Kuroshima, H. (2021). "Socio-Spatial Cognition in Cats: Mentally Mapping Owner's Location from Voice," PLoS ONE 16(11): e0257611. https://doi.org/10.1371/journal.pone.0257611

Tintle, N., Carver, R., Chance, B., Cobb, G., Rossman, A., Roy, S., Swanson, T., and VanderStoep, J. (2019) *Introduction to Statistical Investigations: AP edition*. First edition. Hoboken, New Jersey: John Wiley and Sons.

Tintle, N., Chance, B., Cobb, G., Rossman, A., Roy, S., Swanson, T., and VanderStoep, J. (2020a) *Introduction to Statistical Investigations*. Second edition. Hoboken, New Jersey: John Wiley and Sons.

Tintle, N., Chance, B., McGaughey, K., Roy, S., Swanson, T., and VanderStoep J. (2020b) *Intermediate Statistical Investigations*. First edition. Hoboken, New Jersey: John Wiley and Sons.

Tintle, N., Clark, J., Fischer, K., Chance, B., Cobb, G., Roy, S., Swanson, T., and VanderStoep, J. (2018). "Assessing the Association Between Precourse Metrics of Student Preparation and Student Performance in Introductory Statistics: Results from Early Data on Simulation-Based Inference vs. Nonsimulation-Based Inference," Journal of Statistics Education, 26:2, 103-109, DOI: 10.1080/10691898.2018.1473061

Tintle, N., Rogers, A., Chance, B., Cobb, G., Rossman, A., Roy, S., Swanson, T., VanderStoep, J. (2014). "Quantitative Evidence for the Use of Simulation and Randomization in the Introductory Statistics Course," Proceedings of the 9th International Conference on Teaching Statistics.

Woodard, R., and McGowan, H. (2012). "Redesigning a Large Introductory Course to Incorporate the GAISE Guidelines," Journal of Statistics Education, 20:3, DOI: 10.1080/10691898.2012.1188965