Comparison of Learning Outcomes for Randomization-Based and Traditional Inference Curricula in a Designed Educational Experiment

Karsten Maurer
Iowa State University

Dennis Lock
Iowa State University

Abstract

Conducting inference is a cornerstone upon which the practice of the statistics is based. As such, a large portion of most introductory statistics courses is focused on teaching the fundimentals of statistical inference. The goal of this study is to make a formal comparison of learning outcomes under the traditional and randomization-based inference curricula. A randomized experiment was conducted to administer the two curricula to students in an introductory statistics course. The results indicate that students receiving the randomization-based curriculum have significantly higher learning outcomes for confidence interval related topics. While the results are not comprehensive in assessing the effect on all facets learning, they indicate that learning outcomes for core concepts of statistical inference can be significantly improved with the randomization-based approach.

1. Introduction

Conducting inference is a cornerstone upon which the practice of the statistics is based. As such, a large portion of most introductory statistics courses is focused on teaching the fundimentals of statistical inference. In recent years the approach by which to teach inference in introductory statistics courses has been the topic of growing discussion. The traditional approach to inference curriculum is focused on distributional theory-based methodology, often characterized by use of distributional assumptions, formulas and tables. A modern alternative utilizes a randomization-based approach to the inference curriculum. The randomization-based approach utilizes tactile and computational simulation to run inferencial techniques such as bootstrap confidence intervals and randomization based hypothesis testing. Many proponents of the randomization-based inference curriculum argue that this allows students to be exposed to the core concepts of the inference without first requiring the understanding of theoretical probability distributions.

The focus of the following study is to make a formal comparison of learning outcomes under the traditional and randomization-based inference curricula. The learning outcomes for concepts surrounding inference with confidence intervals and hypothesis testing are of primary concern. A randomized experiment was conducted to administer the two curricula to students in an introductory statistics course. The experimental design allows for causal inference to be drawn about the effect of curriculum type on the learning outcomes. The results indicate that significant improvement in learning outcomes of topics in statistical inference are achieved using the randomization-based teaching methods.

2. Literature Review

With the goal to make proper comparison of traditional versus randomization-based curricula for introductory statistical courses we must first view where each approach stands within the constant evolution of statistics education. Using the term "traditional" to describe the current standard for introductory statistics course curriculum is relative to only the last two decades. Moore chronicled the reform movement of statistic education of the 1980's and 1990's as a period of drastic change in the introductory statistics classroom. The curriculum expanded greatly from a course dominated by theory-based inference methodology to the inclusion of the topics of data exploration, data production, model diagnostics and simulation. The content change indicated a shifting emphasis toward conceptual understanding and applied statistics. Moore also stated, "(w)hat is striking about the current reform movement is not only its momentum but the fact that it centers on pedagogy as much as content" (Moore 1997). The pedagogical push toward active learning combined with the content change and the increasing use of technology to form what may be referred to now as the traditional introductory statistics curriculum.

The tenets of the statistics education reform movement were formalized in the Guidelines for Assessment and Instruction in Statistics Education (GAISE) reports for pre-K-12 and introductory college courses (Aliaga, Cobb, Cuff, Garfield, Gould, Lock, Moore, Rossman, Stephenson, Utts, Velleman, and Witmer 2005). Six recommendations were made in the executive summary of the GAISE report: emphasize statistical literacy and thinking, use real data, stress conceptual over procedural understanding, foster active learning, use technology for both learning and analysis, and use assessment as part of the learning process. In the past decade these principles have been widely adopted in statistics education with a noteworthy increase in technological integration. Technology in the statistics classroom now regularly takes the form of applets, graphing calculators, multimedia materials, and educational, analytical and graphical software (Chance, Ben-Avi, Garfield, and Medina 2007; Rubin 2007). Technological proliferation in the statistics classroom came as a result of technologically receptive statistics educators taking advantage of computation that has become cheaper and more accessible. A large survey of introductory statistics instructors found that 76% of the instructors usually or always require students to use a computer program to explore and analyze data, and 90% of the instructors report a high level of comfort using computer applications to teach introductory statistics (Hassad 2013).

Amidst the drastic increase in the use of technology in introductory statistics education there has been a growing group of educators who believe that the curriculum reform has stopped short of the possibilities that computation can provide. Cobb argues that statistics education has done well to adopt technology to displace tedious calculation but has not effectively changed the approach to teaching inference. Cobb strongly articulates a call for statistics instructors to use randomization-based methods for teaching inference to replace the traditional approach to inference using theory-based methodology. He states, "(o)ur curriculum is needlessly complicated because we put the normal distribution... at the center of our curriculum, instead of the core logic of inference at the center" (Cobb 2007). If we view the introductory statistics course as a constrained optimization problem with statistical literacy and conceptual understanding of inference as the items to maximize, then removing the burden of learning the normal distribution will present the opportunity for more time spent learning core concepts Carver (2011). In recent years, curricula for using a randomization-based approach to inference have been developed by a number of groups of statistics educators (Tintle, Chance, Cobb, Rossman, Roy, Swanson, and Vanderstoep 2014; Lock, Lock, Morgan, Lock, and Lock 2013; CATALST 2012; Carver 2011).

There has been research done on the efficacy of randomization-based inference curricula; however, due to the recency of the curricula development most of this preliminary research has been observational. Budgett, Pfannkuch, Regan & Wild conduct a case study on a small group of students receiving a randomization-based curriculum and found significant learning gains using pre and post testing based on the Comprehensive Assessment of Outcomes in a First Statistics Course (CAOS). This study does not however attempt to make a comparison between the randomization-based approach and traditional approach to teaching inference (Budgett, Pfannkuch, Regan, and Wild 2013). Another pair of studies make comparisons on both learning outcomes and learning retention between the two types of curricula. Tintle, VanderStoep, Holmes, Quisenberry and Swanson found weak evidence for an overall improvement in learning outcomes and significant improvements within the topic of hypothesis testing for the cohort of students receiving the randomization-based curriculum, but the lack of random assignment of student to cohort obstructs the ability to draw any causal conclusions (Tintle, VanderStoep, Holmes, Quisenberry, and Swanson 2011). Tintle, Topliff, Vanderstoep, Holmes and Swanson then found significant evidence for improvements to learning outcome retention after four months for students receiving the randomization-based inference curriculum, but again self-selection of students to cohort prevents establishing a causal link (Tintle, Topliff, VanderStoep, Holmes, and Swanson 2012).

The preliminary research shows promising results for the randomization-based approach to teaching statistical inference. A more rigorous experimental approach to comparing the traditional and randomization-based curricula has been taken in this study in order to establish a causal effect of curriculum on learning outcomes. Section 3 explains the structure and methodology implemented in the educational experiment and the measurement of student learning. Section 4 details the model based approach for assessing the effect of curriculum on specific learning outcomes. Lastly, we discuss the study findings and explore the implications for designing future introductory statistics curricula.

3. Methodology

The subjects for this study were students enrolled in two sections of the Introduction to Statistics, Stat 104, course at Iowa State University in the spring semester of 2014. Stat 104 is an introductory statistics course tailored for students in the agricultural and biological sciences. Of the 112 students to complete the course, 101 students consented to the release of their course data for the purposes of this study. The students who did not consent were treated identically to those who consented, but their data was omitted from the analysis that follows. Students from both sections were randomly assigned to one of the two inference curriculum treatments, creating cohorts A,C and B,D, respectively. Cohorts A and B were explosed to the randomization-based curriculum; while the cohorts C and D were exposed to the traditional curriculum. Student cohorts were the basic units to which room assignments, instruction and curriculum treatments were applied.

The course was administered by the authors in a co-teaching setting for students from all cohorts. The course schedule involved two hours of lecture and two hours of lab per week. The co-teaching strategy was employed as an intentional attribute of the experimental design. The following subsections will detail the curriculum outline for each cohort of students, the experimental design for administering the curricula using the strengths of the co-teaching setup and the data collected for analysis.

3.1. Curricula Structures

To compare the learning outcomes for students receiving the traditional and randomization-based inference curricula we first needed to prepare a curriculum for each approach. Both curricula needed to satisfy the course guidelines set by the Department of Statistics, covering the following topics: univariate and bivariate descriptive statistics, linear regression, experimental design, basic probability rules, the binomial distribution, the normal distribution, sampling distributions, and inference on means and proportions. Each curriculum was composed of lecture, corresponding lecture notes, weekly lab assignments designed for groups of four to five students, weekly homework assignments, a midterm exam and a cumulative final exam. The curricula materials for the course did not involve the use of a textbook, however specific textbooks that roughly follow the structure of each curriculum were recommended as supplementary study materials (Agresti and Franklin 2012; Lock et al. 2013).

Figure 1 outlines how these topics were structured within a weekly schedule for the sixteen week semester for each curriculum. Note that students from all cohorts were exposed to an identical curriculum for all non-inference related topics in the course. This includes identical lecture, course notes, homework assignments, lab assignments and midterm exam during the first half of the semester.

Starting at week 9 the curricula diverge into their respective approaches to inference. Cohorts A and B began the randomization-based inference curriculum in week 9 by first exploring the concepts of sampling distributions then used computer simulation and sampling variability as a basis for exploring inference using bootstrap confidence intervals and randomization tests. Lectures, homework and labs for these cohorts utilized the StatKey software package (Lock et al. 2013) to conduct the randomization-based inference. The randomization-based curriculum then covered normal distributions and how they could be used to conduct inference on means and proportion. While many advocates for randomization-based methods may argue that the normal distribution should be pushed to a second course in statistics, course guidelines required that all students of this introductory statistics course be taught theory-based inference methodology.

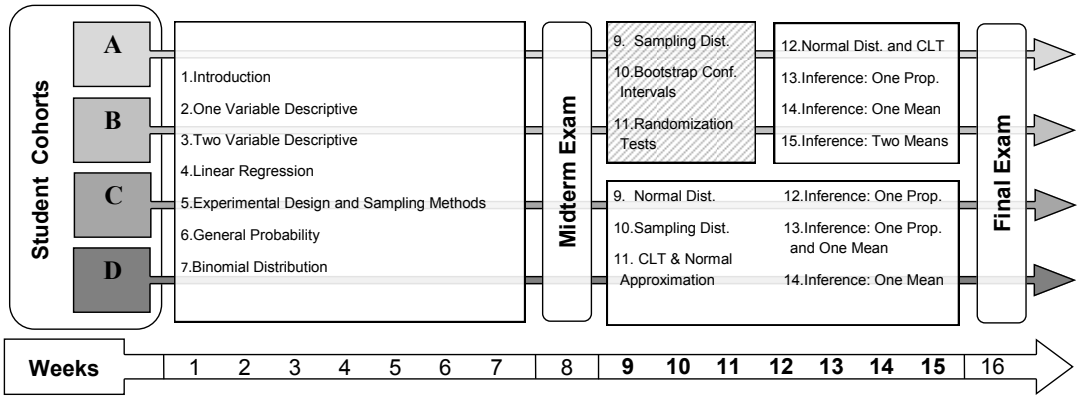

Figure 1: Curricula Schedules

Cohorts C and D progressed through the traditional approach by first learning the normal distribution and use of the normal tables. They were then introduced to applications of the normal approximation within inference. The traditional curriculum utilized simulation to display concepts, but only to the extent of demonstrating that sampling distributions can be approximated by normal distributions under certain conditions.

During the second half of the semester the lectures, course note, homework and lab assignments differed between the two curricula. However, homework and lab assignments were kept similar when they covered similar topics. For example, all cohorts covered the topic of sampling distributions so the lab assignments were nearly identical between the two groups with the exception of a question pertaining to the normal approximation included for the traditional cohorts. By the end of the semester all cohorts covered how to conduct inference using normal theory; however cohorts A and B additionally learned the core concepts of inference using randomization-based methods prior to learning traditional theory-based inference methods.

3.2. Experimental Design

The logistics of administering a course with two distinct curricula and four cohorts of students required a well-structured design and creative scheduling on several fronts. The primary objectives for the experimental design were to eliminate differences in non-inference related curriculum administration to the extent possible, remove the confounding instructor effect on each curriculum and to mitigate the effect of unknown lurking variables through random assignment of students to curricula.

Students were randomly assigned to cohorts during the first week of the course. Of the 101 students who completed the course and consented to the release of their data there were 50 students in the traditional treatment group and 51 students in the randomization-based treatment group. It is also worthwhile to note that of the 4 students to drop the course, all did so prior to week 9; thus, we can safely assume that the inference curriculum treatment did not play a role in the drop. All students who began the inference curricula completed the course.

Students were exposed to identical lecture and lab instruction for the first half of the semester and then diverge into two separate lecture and lab settings for the second half of the semester. This was done to make the experience as similar as possible such that both treatment groups would have the same exposure to terminology and ideas leading up to the inference topics. We could not reassign students to lecture and lab times different than the times which they enrolled, which meant the logistics of the design required preemptive room scheduling and course time scheduling preparations. By working with the department chair and course coordinator before students enrolled into sections we were able to schedule two sections of the course to have identical lecture times but separate lab times. Special room scheduling was required because all students needed to attend the same lecture and lab rooms for the first half of the semester then split into separate lecture and lab rooms after the midterm. This room and course time scheduling allowed for students to be divided into cohorts and attend the lecture or lab specific to their curriculum. The lecture and lab room schedules for each cohort are displayed in Figure 2.

Assigning one instructor to each curriculum would confound the instructor effect and the curriculum effect. To avoid confounding each treatment group would need to receive instruction from both instructors. An alternating weekly schedule was decided upon to spread out the instructor effects on both curricula. A coin was flipped to decide how to match the instructor to the curriculum in order to initialize the alternation. The lecture and lab instruction schedules for each cohort can also be found below in Figure 2. Note that each table has student cohort and times fixed in the column header, reflecting the unchanged time structure that each student enrolled into. The instructor and room location change throughout the rows as they did throughout the course.

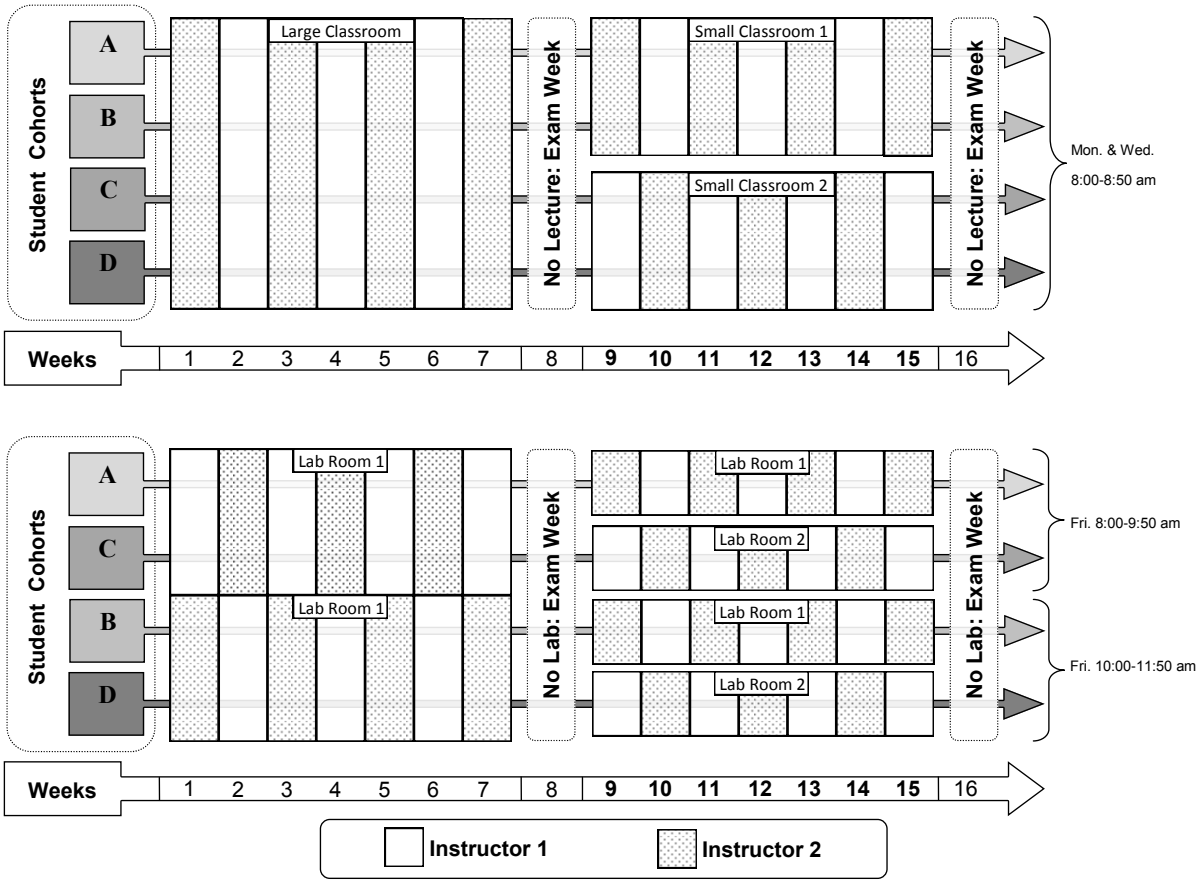

Figure 2: Instructors Schedule

3.3. Data Collection

In order to measure learning outcomes for specific inference concepts we utilized question sets from the Assessment Resources Tools for Improving Statistical Thinking (ARTIST) for the topics of confidence intervals and hypothesis testing (ARTIST 2006). The ARTIST scaled question sets each consist of 10 multiple choice questions that are geared toward critical thinking about each inference topic. These questions were administered as part of the written final exam for all students on the same day and time. The ARTIST scaled scores for the topics of confidence interval and hypothesis testing were recorded for each student. The multiple choice questions for the ARTIST scaled topics can be found in Appendix A.

The final exam also included two problems that tested the student's ability to conduct statistical inference in an applied setting using theory-based methodology. Each problem was based on a hypothetical scenario where data has been collected and inference needs to be conducted using the traditional approach using formulas and tables. The first problem provided data summaries and students needed to construct and interpret a confidence interval for a single population mean. The second problem required students to conduct a hypothesis test for a single proportion based on another set of data summaries. The applied inference problem scores for each student are not used for the primary analysis on learning outcomes but are included for an interesting peripheral analysis on student ability to conduct inference using traditional theory-based methods. The problems and grading rubics can be found in Appendix A.

In addition to the ARTIST scaled question set scores, data was collected from the first eight weeks of the course – prior to student exposure to an inference curriculum. We have scores from homeworks 1 to 7, lab assignments 1 to 7 and the midterm exam for each student. Since all of these items were administered and graded simultaniously for all students before being assigned to a curricula, the scores from these weeks will be referred to as the "pre-treatment measurements". Lastly, the data include the cohort to which each student belonged.

The research proposal approved by the Institutional Review Board specified that students' data would be entirely deidentified following the course, including all demographic infomation. At the conclusion of the semester the data for the 101 students who consented to the release of their data were saved, with names and identity information removed, to a spreadsheet. The deidentified student data was imported to `R` for the analysis described in the sections below.

4. Analysis

The primary goal of the analysis is to investigate if there is a curricula effect on inference concept learning outcomes. Our data includes ARTIST scaled topic scores for confidence intervals and hypothesis tests which we use as the responses for the comparison of curricula. A model based approach is used to assess curricula effect while controlling for pre-treatment differences between students. With two dimensional response variables and an assortment of covariates we employ a multivariate analysis of covariance (MANCOVA) model.

Both curricula groups were required to learn to conduct normal-based inference. This leads to another question of interest. Does the added randomization-based material turn out to be detrimental to student's ability to use distributional theory-based methods to conduct inference? Two applied inference questions were included on the final exam that required students to use of theoretical methods to conduct inference. These applied questions were used as the responses in a separate MANCOVA model to check for a treatment effect. Model parameterization used for these Bivariate MANCOVA models can be found in Appendix B.

4.1. Modeling ARTIST Outcomes

We begin with the model for the ARTIST scaled topic scores. Many of the pre-treatment variables are highly correlated. To select a model with only the most predictive pre-treatment covariates, model selection was conducted by first running backward selection based on AIC then removing further covariates that posed collinearity issues. The model selected for final analysis included three covariates: an indicator variable for the curriculum treatment group, lab 5 score and the midterm score. The midterm tested students on materials from weeks 1-7 and lab 5 assess understanding of topics related to random selection techniques. We will refer to this selected model as the "ARTIST Model".

Model fit for the ARTIST Model was assessed to be satisfactory. Residual plots indicate that the assumptions of linearity and homoskedasticity hold, see Appendix C. The response variables display approximate univariate and bivariate normality. MANOVA models are best behaved with moderate correlation between the response variables. The correlation between the ARTIST scaled topic scores for confidence intervals and hypothesis tests is acceptable with, $r = 0.288$. Lastly, the assumption of independence between student response scores is justified because the scores were achieved through students working individually in a controlled testing environment.

To test for overall covariate effects on the multivariate responses we use Pillai's Λ. Table 1 shows a weak overall effect of the curriculum treatment on the ARTIST scaled topic scores and the treatment has a weakly significant overall effect. Pillai's Λ measures the overall effect by combining the effects on the bivariate responses. This prompts us to investigate the treatment effect on the ARTIST scaled topic scores for confidence intervals and hypothesis tests separately, to see if the weak overall effect is driven by a significant effect on one of the two scores.

	Pillai's Λ	Approx. F	Pr(>|F|)
Midterm	0.2109	12.8277	0.0000
Lab 5	0.0792	4.1261	0.0191
Treatment	0.0469	2.3605	0.0998

Table 1: Tests for overall covariate effects on ARTIST question scores using Pillai's Λ.

To investigate the effect of the curriculum treatment on each ARTIST scaled topic score, we analyze these two linear models that comprise the overall MANCOVA model. Table 2 displays the coeffients of the linear model fit to the ARTIST scaled score for confidence interval learning outcomes along with covariate ranges to provide context to coefficent magnitudes. We find that midterm, lab 5 and the curriculum treatment effect are significant. Specifically, the randomization-based inference group scored significantly higher by 0.7149 out of a possible 10 points, a 7.146% improvement in confidence interval learning outcomes on the ARTIST scale while controlling for midterm and lab 5 scores.

	Covariate Values	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1	1.4648	1.0135	1.45	0.1516
Midterm	{0,1,...100}	0.0477	0.0113	4.22	0.0001
Lab 5	{0,1,...100}	0.01827	0.6337	2.88	0.0048
Treatment	{0,1}	0.7146	0.3382	2.11	0.0371

Table 2: Linear model for ARTIST scaled score – out of 10 points – for confidence interval topic.

Table 3 displays the coeffients of the linear model fit to the ARTIST scaled score for hypothesis test learning outcomes. We find that only the midterm score is significant for predicting learning outcomes for hypothesis testing. There was no statistically significant curriculum treatment effect.

	Covariate Values	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1	2.1053	1.0584	1.99	0.0495
Midterm	{0,1,...100}	0.0386	0.0118	3.27	0.0015
Lab 5	{0,1,...100}	0.0085	0.6618	1.29	0.1996
Treatment	{0,1}	0.3050	0.3532	0.86	0.3900

Table 3: Linear model for ARTIST scaled score – out of 10 points – for hypothesis testing topic.

4.2. Modeling Applied Theory-Based Inference Scores

As with the MANCOVA model for ARTIST scaled question scores, we consider all pre-treatment measurements in a new model for the two applied theory-based inference questions. Backward stepwise selection was used to obtain a reduced MANCOVA model in a process very similar to that described in Subsection 4.1. We will refer to the selected model here as the "Applied Model".

Table 4 shows that the Pillai's Λ for the treatment term in the Applied Model indicates that there was no overall effect of curriculum treatment on the scores for the applied inference problems. This is of particular interest because students receiving the randomization-based curriculum had three weeks less of coursework involving the use of normal distributions and tables; however, it did not lead to a significant difference in the ability to properly apply theory-based methods. Residual plots for the Applied Model may be found in Appendix C.

	Pillai's Λ	Approx. F	Pr(>|F|)
Midterm	0.3238	22.9881	0.0000
Homework 2	0.0870	4.5750	0.0127
Treatment	0.0108	0.5217	0.5952

Table 4: Tests for overall covariate effects on Applied question scores using Pillai's Λ.

5. Discussion and Conclusions

The results indicate that students receiving the randomization-based curriculum have significantly higher learning outcomes for confidence interval topics. The magnitiude of the improvement was 7.146% on the ARTIST scale. There was no significant difference between traditional and randomization-based curricula on hypothesis test topic learning outcomes. Finding a significant positive effect of the randomization-based curriculum on learning outcomes for confidence interval topics raises questions of representation and causality.

We must bear in mind the population for which these results may be representative. The study was conducted with undergraduate students enrolled in an introductory statistics course at a large public midwestern university. The course is required for students in the agricultural and biological science. Students in the course are predominantely sophomores and juniors. The results are only applicable to the extent to which these 101 students represent the broader population of introductory statistics students.

The experimental design bolsters the establishment of a causal effect through utilization of control of non-inferencial course components and random assignment of students to treatment groups. There are two assumptions that we must make to justify a claim of causality. We must assume that the random assignment successfully eliminated all possible lurking variables; however this is the assumption made by all randomized experiments. We must make the assumption that the instructor effect on learning outcomes has been elimiated by the weekly alternation of instructors. We beleive both assumptions are justifiable due to the care taken with randomization and instruction alternation.

An issue that is more problematic than the assumptions made about causality is that the treatment itself was a half semester curriculum – a highly complex combination of lesson plans, lecture content, assignments and technology use. The treatment complexity poses a problem in identifying precisely what component of the curriculum caused the improvement in confidence interval related learning outcomes.

It is important to recall that due to departmental requirements for the course, the treatment group learned randomization-based inference *in addition to* an abbreviated unit on theory-based inference methods. The randomization-based curriculum lead to no significant difference in learning outcomes for hypothesis testing on the ARTIST scale, dispite the added complexity of learing additional concepts for conducting randomization tests. In addition, the added complexity related to bootstrapping confidence intervals actually improved learning outcomes for confidence intervals on the ARTIST scale.

The analysis of the Applied Model found no significant difference between curriculum groups for scores on the applied theory-based inference problems. This implies that dispite the increased complexity of the randomization-based material and the shortened exposure to theory-based inference concepts, there was no detriment to students' performance in conducting inference using theory-based methods. It should be noted that the applied questions from the final exam were written by the authors and have not been assessed as reliable metrics for learning outcomes. Thus, the results are reported as supplementary to the discussion on learning outcomes measured by the ARTIST scaled topics.

The randomization-based approach to inference that we employed achieved an improvement in measured learning outcomes related to confidence intervals, but there are many possible ways to implement randomization-based inference within a course. One noteworthy characteristic was that the randomization-based curriculum that was employed in this study utilized bootstrapping to teach the concepts of confidence intervals as opposed to inverting a randomization test. This study does not attempt to imply that all implementations of the randomization-based approach would achieve that same improvement in learning outcomes.

We believe that the results of this study are affirming to efficacy of randomization-based methods for teaching inference. We found that the randomization-based curriculum lead to significant improvement in the learning outcomes associated with confidence intervals but no significant difference from the traditional approach for learning outcomes associated with hypothesis testing. While these results are clearly not comprehensive in assessing the effect of a randomization-based curriculum on all facets learning, they indicate that learning outcomes for a core concept of statistical inference can be significantly improved with the approach.

A. Appendix of Exam Questions

A.1. ARTIST Scaled Multiple Choice Question Set for Confidence Intervals

1. Answer the following general multiple choice questions regarding confidence intervals. There is only one correct answer for each (circle the best option).

i. Two different samples will be taken from the same population of test scores where the population mean and standard deviation are unknown. The first sample will have 25 data values, and the second sample will have 64 data values. A 95% confidence interval will be constructed for each sample to estimate the population mean. Which confidence interval would you expect to have greater precision (a smaller width) for estimating the population mean?

a. I expect the confidence interval based on the sample of 64 data values to be more precise.
b. I expect both confidence intervals to have the same precision.
c. I expect the confidence interval based on the sample of 25 data values to be more precise.

ii. A 95% confidence interval is computed to estimate the mean household income for a city. Which of the following values will definitely be within the limits of this confidence interval?

a. The population mean
b. The sample mean
c. The standard deviation of the sample mean
d. None of the above

iii. Each of the 110 students in a statistics class selects a different random sample of 35 Quiz scores from a population of 5000 scores they are given. Using their data, each student constructs a 90% confidence interval for the average Quiz score of the 5000 students. Which of the following conclusions is correct?

a. About 10% of the sample means will not be included in the confidence intervals.
b. About 90% of the confidence intervals will contain .
c. It is probable that 90% of the confidence intervals will be identical.
d. About 10% of the raw scores in the samples will not be found in these confidence intervals

iv. A 95% confidence interval for the mean reading achievement score for a population of third grade students is (43, 49). The margin of error of this interval is:

a. 5
b. 3
c. 6

v. Justin and Hayley conducted a mission to a new planet, Planet X, to study arm length. They took a random sample of 100 Planet X residents and calculated a 95% confidence interval for the mean arm length. What does a 95% confidence interval for arm length tell us in this

case? Select the best answer:

a. I am 95% confident that this interval includes the sample mean (x) arm length.
b. I am confident that most (95%) of all Planet X residents will have an arm length within this interval.
c. I am 95% confident that most Planet X residents will have arm lengths within this interval.
d. I am 95% confident that this interval includes the population mean arm length.

vi. Suppose that a random sample of 41 state college students is asked to measure the length of their right foot in centimeters. A 95% confidence interval for the mean foot length for students at this university turns out to be (21.709, 25.091). If instead a 90% confidence interval was calculated, how would it differ from the 95% confidence interval?

a. The 90% confidence interval would be narrower.
b. The 90% confidence interval would be wider.
c. The 90% confidence interval would be the same as the 95% confidence interval.

vii. A pollster took a random sample of 100 students from a large university and computed a confidence interval to estimate the percentage of students who were planning to vote in the upcoming election. The pollster felt that the confidence interval was too wide to provide a precise estimate of the population parameter. What could the pollster have done to produce a narrower confidence interval that would produce a more precise estimate of the percentage of all university students who plan to vote in the upcoming election?

a. Increase the sample size to 150.
b. Increase the confidence level to 99%.
c. Both a and b
d. None of the above

viii. A newspaper article states with 95% confidence that 55% to 65% of all high school students in the United States claim that they could get a hand gun if they wanted one. This confidence interval is based on a poll of 2000 high school students in Detroit. How would you interpret the confidence interval from this newspaper article?

a. 95% of large urban cities in the United States have 55% to 65% high school students who could get a hand gun.
b. If we took many samples of high school students from different urban cities, 95% of the samples would have between 55% and 65% high school students who could get hand guns.
c. You cannot use this confidence interval to generalize to all teenagers in the United States because of the way the sample was taken.
d. We can be 95% confident that between 55% and 65% of all United States high school students could get a hand gun.

ix. The Gallup poll (August 23, 2002) reported that 53% of Americans said they would favor sending American ground troops to the Persian Gulf area in an attempt to remove Hussein from power. The poll also reported that the margin of error for this poll was 4%. What does the margin of error of 4% indicate?

a. There is a 4% chance that the estimate of 53% is wrong.
b. The percent of Americans who are in favor is probably higher than 53% and closer to 57%.
c. The percent of Americans who are in favor is estimated to be between 49% and 57%.

x. Suppose two researchers want to estimate the proportion of American college students who favor abolishing the penny. They both want to have about the same margin of error to estimate this proportion. However, Researcher 1 wants to estimate with 99% confidence and Researcher 2 wants to estimate with 95% confidence. Which researcher would need more students for her study in order to obtain the desired margin of error?

a. Researcher 1.
b. Researcher 2.
c. Both researchers would need the same number of subjects.
d. It is impossible to obtain the same margin of error with the two different confidence levels.

A.2. ARTIST Scaled Multiple Choice Question Set for Hypothesis Testing

2. Answer the following general multiple choice questions regarding hypothesis testing. There is only one correct answer for each (circle the best option). As a helpful note the term "statistically significant" means that you reject the null hypothesis.

i. The makers of Mini-Oats cereal have an automated packaging machine that is set to fill boxes with 24 ounces of cereal. At various times in the packaging process, a random sample of 100 boxes is taken to see if the machine is filling the boxes with an average of 24 ounces of cereal. Which of the following is a statement of the null hypothesis being tested?

a. The machine is filling the boxes with the proper amount of cereal.
b. The machine is not filling the boxes with the proper amount of cereal.
c. The machine is not putting enough cereal in the boxes.

ii. A research article gives a p-value of .001 in the analysis section. Which definition of a p-value is the most accurate?

a. the probability that the observed outcome will occur again.
b. the probability of observing an outcome as extreme or more extreme than the one observed if the null hypothesis is true.
c. the value that an observed outcome must reach in order to be considered significant under the null hypothesis.
d. the probability that the null hypothesis is true.

iii. If a researcher was hoping to show that the results of an experiment were statistically significant they would prefer:

a. a large p-value
b. a small p-value
c. p-values are not related to statistical significance

iv. A researcher compares men and women on 100 different variables using a difference in means t-test. He sets the level of significance at 0.05 and then carries out 100 independent t-tests (one for each variable) on these data. If, for each test, the null hypothesis is actually true, about how many "statistically significant" results will be produced?

a. 0
b. 5
c. 10
d. none of the above

Problems (v) and (vi) refer to the following situation: Food inspectors inspect samples of food products to see if they are safe. This can be thought of as a hypothesis test where Null: the food is safe, and Alternative: the food is not safe. Identify each of the following statements as a Type I or a Type II error.

v. The inspector says the food is safe but it actually is not safe.

a. Type I
b. Type II

vi. The inspector says the food is not safe but is actually safe.

a. Type I
b. Type II

vii. A newspaper article claims that the average age for people who receive food stamps is 40 years. You believe that the average age is less than that. You take a random sample of 100 people who receive food stamps, and find their average age to be 39.2 years. You find that this is significantly lower than the age of 40 stated in the article (p-value $< .05$). What would be an appropriate interpretation of this result?

a. The statistically significant result indicates that the majority of people who receive food stamps is younger than 40.
b. Although the result is statistically significant, the difference in age is not of practical importance.
c. An error must have been made. This difference is too small to be statistically significant.

viii. A newspaper article stated that the US Supreme Court received 812 letters from around the country on the subject of whether to ban cameras from the courtroom. Of these 812 letters, 800 expressed the opinion that cameras should be banned. A statistics student was going to use this sample information to conduct a test of significance of whether more than 95% of all American adults feel that cameras should be banned from the courtroom. What would you tell this student?

a. This is a large enough sample to provide an accurate estimate of the American public's

opinion on the issue.
b. The necessary conditions for a test of significance are not satisfied, so no statistical test should be performed.
c. With such a large number of people favoring the notion that cameras be banned, there is no need for a statistical test.

ix. A researcher conducts an experiment on human memory and recruits 15 people to participate in her study. She performs the experiment and analyzes the results. She obtains a p-value of .17. Which of the following is a reasonable interpretation of her results?

a. This proves that her experimental treatment has no effect on memory.
b. There could be a treatment effect, but the sample size was too small to detect it.
c. She should reject the null hypothesis.
d. There is evidence of a small effect on memory by her experimental treatment.

x. It is reported that scores on a particular test of historical trivia given to high school students are approximately normally distributed with a mean of 85. Mrs. Rose believes that her 5 classes of high school seniors will score significantly better than the national average on this test. At the end of the semester, Mrs. Rose administers the historical trivia test to her students. The students score an average of 89 on this test. After conducting the appropriate statistical test, Mrs. Rose finds that the p-value is .0025. Which of the following is the best interpretation of the p-value?

a. A p-value of .0025 provides strong evidence that Mrs. Rose's class outperformed high school students across the nation.
b. A p-value of .0025 indicates that there is a very small chance that Mrs. Rose's class outperformed high school students across the nation.
c. A p-value of .0025 provides evidence that Mrs. Rose is an exceptional teacher who was able to prepare her students well for this national test.
d. None of the above.

A.3. Applied Theory-Based Confidence Interval Question

3. Farmer Cindy is in charge of the chickens on her family's farm, and is curious about the average number of eggs the entire flock produces in a month. Observing the entire flock would be time consuming, so she approaches you asking what her options are. You inform her that 30 chickens should be selected at random to be observed for a month. After the month she observes the average number of eggs her sample of 30 chickens produced is 25 eggs with a standard deviation of 6 eggs.

1. (5 pts) Construct a 95% confidence interval for the average number of eggs chickens from her entire population produce in a month. (You don't need to check any conditions here)
2. (3 pts) Interpret the 95% confidence interval constructed in the previous part:
3. (2 pts) Cindy is disappointed with the width of the interval you provide for her, suggest to her two ways she could obtain a narrower confidence interval.

A.4. Applied Theory-Based Hypothesis Testing Question

4. Cindy becomes concerned about a disease some of her chickens are catching that causes a decrease in the chicken's egg production. She is interested in the proportion of her entire flock that has the disease, but detecting the disease requires taking blood from the chicken which is expensive and time consuming. After working with you in the past, she understands that she can estimate this proportion by taking just a sample of her chickens! Assume she selects a sample of 100 chickens in the best possible way and observed that 15 of the chickens had the disease.

Cindy's pessimistic guess is that 25% of the flock has the disease. Complete the following steps to test if the proportion of her entire population diseased is less than 0.25.

1. (2 pts) State the Null and Alternative hypothesis:
2. (2 pts) Check the conditions for a hypothesis test:
3. (2 pts) Calculate the test statistic:
4. (2 pts) Find the p-value:
5. (3 pts) Make a decision about your hypothesis and state your conclusion in context of the problem.

B. Appendix on Bivariate MANCOVA Model Specifications

$$\mathbf{Y_{n\times 2}} = \mathbf{X_{n\times(P+2)}B_{(P+2)\times 2}} + \epsilon_{\mathbf{n\times 2}} \sim \mathbf{MVN(XB, \Sigma)}$$

$$y_{ik} = \tau_k \mathbb{1}_{\{i\in T\}} + \beta_{0k} + \left(\sum_{p=1}^{P} x_{ip}\beta_{pk}\right) + \epsilon_{ik}, \text{ where}$$

$$\begin{gathered}\text{student } i \in \{1, 2, ..., n\},\\ \text{response } k \in \{1, 2\},\\ \text{covariate } p \in \{1, 2, ..., P\}.\end{gathered}$$

Term	Definition
y_{ik}	k^{th} response from i^{th} student.
τ_k	Treatment group (randomization-based class) effect
$\mathbb{1}_{\{i\in T\}}$	Indicator function that student i is in treatment group
β_{0k}	Common intercept for predicting k^{th} response
x_{ip}	p^{th} covariate score from i^{th} student.
β_{kp}	p^{th} model coefficient for predicting k^{th} response
ϵ_{ik}	k^{th} error term from i^{th} student.

C. Appendix of MANCOVA Residual Plots

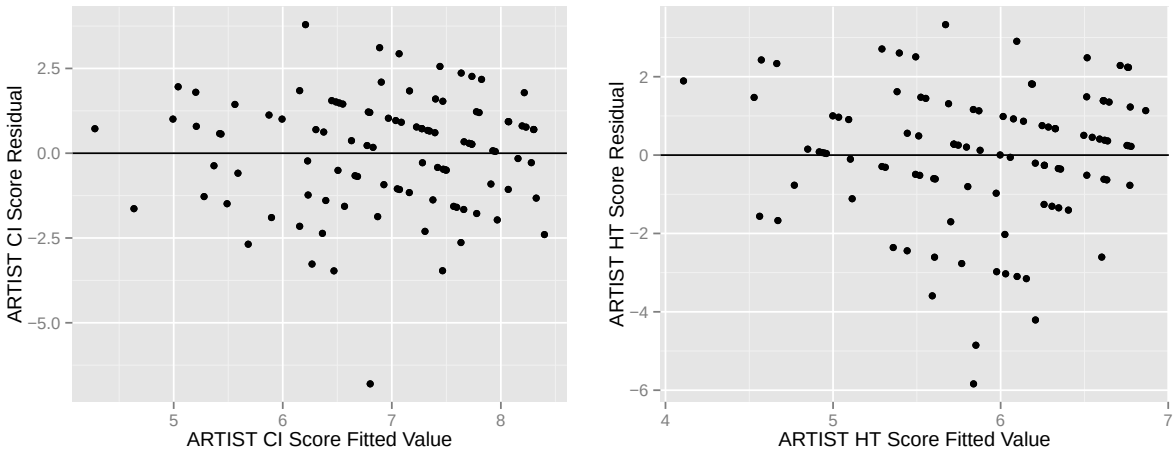

Figure 3: ARTIST Model Residual Plots.

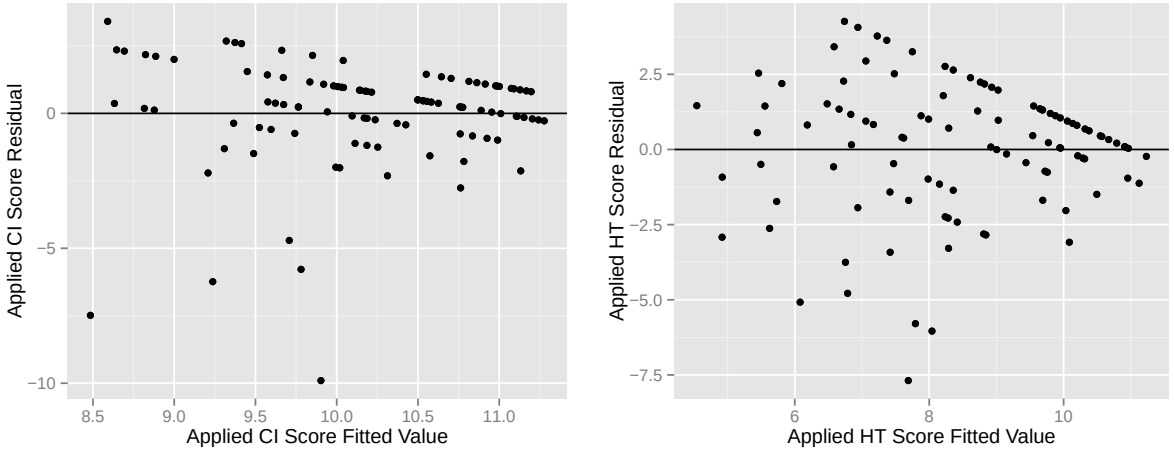

Figure 4: Applied Model Residual Plots.

References

Agresti A, Franklin C (2012). Statistics: The Art and Science of Learning From Data. Pearson, Upper Saddle River, NJ. ISBN 0321756320.

Aliaga M, Cobb G, Cuff C, Garfield J, Gould R, Lock R, Moore T, Rossman A, Stephenson B, Utts J, Velleman P, Witmer J (2005). "Guidelines for Assessment and Instruction in Statistics Education: College Report."

ARTIST (2006). apps3.cehd.umn.edu/artist/tests/index.html. Accessed: 09/08/2014.

Budgett S, Pfannkuch M, Regan M, Wild C (2013). "Dynamic Visualisation and the Randomization Test." Technology Innovations in Statistics Education.

Carver R (2011). "Introductory Statistics Unconstrained by Computability: A New Cobb Salad." Technology Innovations in Statistics Education.

CATALST (2012). "Change Agents for Teaching and Learning Statistics (CATALST) Materials." www.tc.umn.edu/~catalst/materials. Accessed: 09/06/2014.

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

Cobb G (2007). "The Introductory Statistics Course: A Ptolemaic Curriculum." Technology Innovations in Statistics Education.

Hassad RA (2013). "Faculty Attitude towards Technology-Assisted Instruction for Introductory Statistics in the Context of Educational Reform." Technology Innovations in Statistics Education.

Lock R, Lock P, Morgan K, Lock E, Lock D (2013). Statistics: Unlocking the Power of Data. Wiley, Hoboken, NJ. ISBN 9780470601877.

Moore DS (1997). "New Pedagogy and New Content: The Case of Statistics." International Statistical Review.

Rubin A (2007). "Much Has Changed; Little Has Changed: Revisiting the Role of Technology in Statistics Education 1992-2007." Technology Innovations in Statistics Education.

Tintle N, Chance B, Cobb G, Rossman A, Roy S, Swanson T, Vanderstoep J (2014). Introduction to Statistical Investigation. Wiley, Hoboken, NJ. ISBN 1118956672.

Tintle N, Topliff K, VanderStoep J, Holmes V, Swanson T (2012). "Retention of Statistical Concepts in a Preliminary Randomization-Based Introductory Statistics Curriculum." Statistics Education Research Journal.

Tintle N, VanderStoep J, Holmes V, Quisenberry B, Swanson T (2011). "Development and Assessment of Preliminary Randomization-Based Introductory Statistics Curriculum." Journal of Statistics Education.