Expanding the Toolkit: The Potential for Bayesian Methods in Education Research (Symposium)

Symposium Abstract, SREE 2017 Spring Conference

Organizer: Alexandra Resch, aresch@mathematica-mpr.com

Symposium Title: Expanding the Toolkit: The Potential for Bayesian Methods in Education Research

Conference section: Research Methods, Research to Practice

Symposium Justification

Bayesian statistical methods have become more feasible to implement with advances in computing but are not commonly used in educational research. In contrast to frequentist approaches that take hypotheses (and the associated parameters) as fixed, Bayesian methods take data as fixed and hypotheses as random. This difference means that Bayesian inference can take the form of an intuitive probabilistic statement about the likelihood of a particular hypothesis being true. Frequentist results are sometimes framed this way, but this framing is incorrect and can be very misleading. Bayesian methods also allow the incorporation of prior information and can facilitate the systematic combination of evidence from multiple sources.

These features make the Bayesian approach particularly well-suited to informing ground-level educational decisions. Educators and school leaders often need to make decisions without conclusive evidence but do want to use evidence to inform their decisions. The ability of Bayesian methods to intuitively characterize the uncertainty involved in any particular analysis can make research more relevant and useful to education decision-makers.

The papers in this symposium explore the potential for using Bayesian methods in education research and decision-making. The first paper provides an introduction to Bayesian methods and some examples of how they are used outside of education research. The second paper presents results from an experimental study of how people interpret and use the results of a study of an educational technology when the results are presented with frequentist versus Bayesian framing. The third paper presents an application of simple Bayesian analyses to real-world education technology evaluations. The fourth paper presents a more complex application of Bayesian analyses, using a simulation study to demonstrate that a Bayesian adaptive design can provide better inference with smaller samples.

Paper Titles with Presenting Authors and Affiliations:

Paper 1 – Why Bother With Bayes?, Thomas Louis, Johns Hopkins Bloomberg School of Public Health, tlouis@jhu.edu

Paper 2 - Comparing Bayesian and Frequentist Inference for Decision-Making, Ignacio Martinez, Mathematica Policy Research, IMartinez@Mathematica-Mpr.com

Paper 3 - Simple Application of Bayesian Methods for School-Level Decisions, Alexandra Resch, Mathematica Policy Research, AResch@Mathematica-Mpr.com

Paper 4 - What Works for Whom? A Bayesian Approach to Channeling Big Data Streams for Policy Analysis, Jonathan Gellar, Mathematica Policy Research, JGellar@Mathematica-Mpr.com

Discussant: Elizabeth Stuart, Johns Hopkins University, estuart@jhu.edu

Why Bother With Bayes?

Thomas Louis

The use of Bayesian-based designs and analyses in biomedical, environmental, educational, policy and many other contexts has burgeoned, even though its use entails additional overhead. Consequently, it is evident that statisticians and collaborators are increasingly finding the approach worth the bother. In the context of this escalating incidence, I highlight a subset of the potential advantages of the formalism in study design ("Everyone is a Bayesian in the design phase"), conduct, analysis and reporting. Approaches include designs and analyses with required frequentist properties (Bayes for frequentist) and for fully Bayesian goals (Bayes for Bayes). Examples are drawn from sample size estimation, design and analysis of cluster randomized studies, use of historical controls, frequentist CI coverage, evaluating subgroups, dealing with multiplicity, ranking and other nonstandard goals.

The Bayesian approach is by no means a panacea. Valid development and application places additional obligations on the investigative team, and so it isn't always worth the effort. However, the investment can pay big dividends, the cost/benefit relation is increasingly attractive, and in many situations it is definitely worth the bother.

Comparing Bayesian and Frequentist Inference for Decision-Making

Ignacio Martinez*, Alexandra Resch and Mariel McKenzie Finucane

Decision makers, like school district administrators and principals, want to use data to inform their decisions. For example, a school principal may want to use data to choose whether or not to invest in a new technology that promises to help improve student achievement. Most of the research that is available for their consumption uses the null hypothesis significance testing (NHST) paradigm to determine if there is evidence that the new technology is better than business as usual.

Under the NHST approach, also known as the frequentist approach, the researcher typically formulates a "null hypothesis" stating that the intervention has no effect on achievement, and computes a number known as p-value to determine whether to accept or reject the null hypothesis. The standard practice is to reject your null hypothesis if the p-value is below 0.05. On the other hand, if the p-value is higher than 0.05, the standard practice is to say that you cannot reject the null hypothesis. In this case we say that the observed differences are not statistically significant and cannot say that the intervention works with a high level of confidence. Notice that this is not the same as saying that the intervention you are testing does not work. This p-value is a function of the data and the null hypothesis; often large sample sizes are needed in order to meet this threshold for rigor. Alas, the decision makers we have in mind often have a small sample size problem making the 0.05 threshold very difficult to meet.

P-values are hard to understand. As typically used, they quantify the probability of observing results as extreme as –or more extreme than—those observed, if the unobserved true effect were zero (the null hypothesis). In March 2016, the American Statistical Association (ASA) released a statement to address the problem that p-values are constantly misinterpreted (Wasserstein & Lazar 2016). This statement enumerates the following six principles:

1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Decisions have, and will, be made even if a researcher cannot reject that the intervention had no effect with a 95% confidence level. Additionally, this de facto 0.05 threshold is arbitrary and it has long been argued that this threshold should depend on the specific application.

Moreover, McShane and Gal (2015) show that researchers across a variety of field make erroneous statements and judgments when presented with evidence that fails to attain statistical significance. For example, the subjects of that study were much more likely to correctly summarize the results if the p-value was set 0.01 than to 0.27.

The Bayesian paradigm is an alternative to this frequentist approach. The Bayesian paradigm takes data as fixed and estimates the likelihood that hypotheses are true. This approach allows intuitive probability statements in plain English to report findings. For example, a Bayesian approach supports a statement that there is an X percent chance that the new technology improved the outcome of interest by at least Y percent. Therefore, we believe that reporting findings using this paradigm can be more useful for decision makers. Moreover, this framing could result in different decisions being made and could affect the confidence the decision maker has about his or her choices.

In this paper we will assess whether decision makes make different choices when presented information under the frequentist or Bayesian paradigms. We also assess whether participants are more confident that they made the right choice under one of the two paradigms. To answer these questions, we are conducting an online experiment where participants are presented with information are asked a series of questions about whether or not they would invest in a new technology that promises to improve student achievement. The information is a 1-page executive summary of a hypothetical study of the educational technology using either frequentist or Bayesian framing. This experiment is currently underway.

Citations:

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*.

McShane, B. B., & Gal, D. (2015). Blinding us to the obvious? The effect of statistical training on the evaluation of evidence. *Management Science*,*62*(6), 1707-1718.

Simple Application of Bayesian Methods for School-Level Decisions

Alexandra Resch* and Ignacio Martinez

Schools, district administrators, and teachers are constantly making decisions that have the potential to affect students' outcomes. The purchase and use of educational software applications, or "apps," is one area in which educators regularly make and revisit decisions. Currently, objective scientific evidence does not play an important role in this decision making process. Instead, experience, subjective recommendations, and marketing materials are often the only information sources available to support these choices. This is partly due to a general perception that rigorous research has to be complicated, expensive, and take a long time. Educators and school administrators also express frustration with the mismatch between the desire to pilot a product in a small number of schools or classrooms and the need for large sample sizes to have sufficient statistical power for traditional analyses.

The Office of Educational Technology at the US Department of Education has sponsored the development of tools to meet the needs of districts seeking to evaluate the education technology products they use. The free online toolkit includes a series of tools to support districts at all stages of the research process, from considering what questions are worth pursuing to planning and conducting rigorous evaluations and interpreting the results. The ultimate goal is to enable educators to generate timely evidence on the effectiveness of these products, along with strategies for implementing them, for their own students and schools. Schools and districts can use the knowledge and evidence generated by this process ensure that all students have access to effective programs and educational opportunities.

This toolkit uses a simple Bayesian approach for the impact analysis. School administrators are making local decisions about whether to continue with a specific technology product. They are seeking to answer a question like: What's the likelihood we're better off using this product in our schools than we would be otherwise? This question is well suited to the Bayesian approach and the results can accurately be presented in the form of a probability that the product is better than an alternative. The toolkit's analysis uses uninformative (or flat) priors, so does not ask the user to input subjective information that may drive the results. Further expansions of the toolkit may incorporate informative priors to take advantage of results of previous studies, particularly for subsequent studies of the same technology.

This paper will introduce the specific Bayesian analysis employed in the toolkit and will present several illustrative examples of how the results were used in evaluations with districts piloting the rapid cycle evaluation toolkit.

What Works for Whom? A Bayesian Approach to Channeling Big Data Streams for Policy Analysis

Mariel McKenzie Finucane, Ignacio Martinez, and Scott Cody

Presenter: Jonathan Gellar

Schools routinely use different instructional approaches and products with different groups of students, attempting to match educational materials and techniques to student needs. These decisions are often made with little rigorous evidence to support educators' beliefs about which products work best for whom. At the same time, educational technologies are collecting vast amounts of data that could be used to rigorously inform these decisions. Bayesian approaches to designing RCTs can help collect data more efficiently. A Bayesian adaptive design adapts to accumulating evidence: over the course of the trial, more study subjects are allocated to treatment arms that are more promising, given the specific subgroup that each subject comes from. This approach, which is based on the design of two recent clinical drug trials (Barker et al., 2009; Kim et al., 2011), provides valid estimates of heterogeneous causal effects sooner and with smaller sample sizes than would be required in a traditional RCT. To our knowledge, this strategy has not yet been applied to education or social policy research.

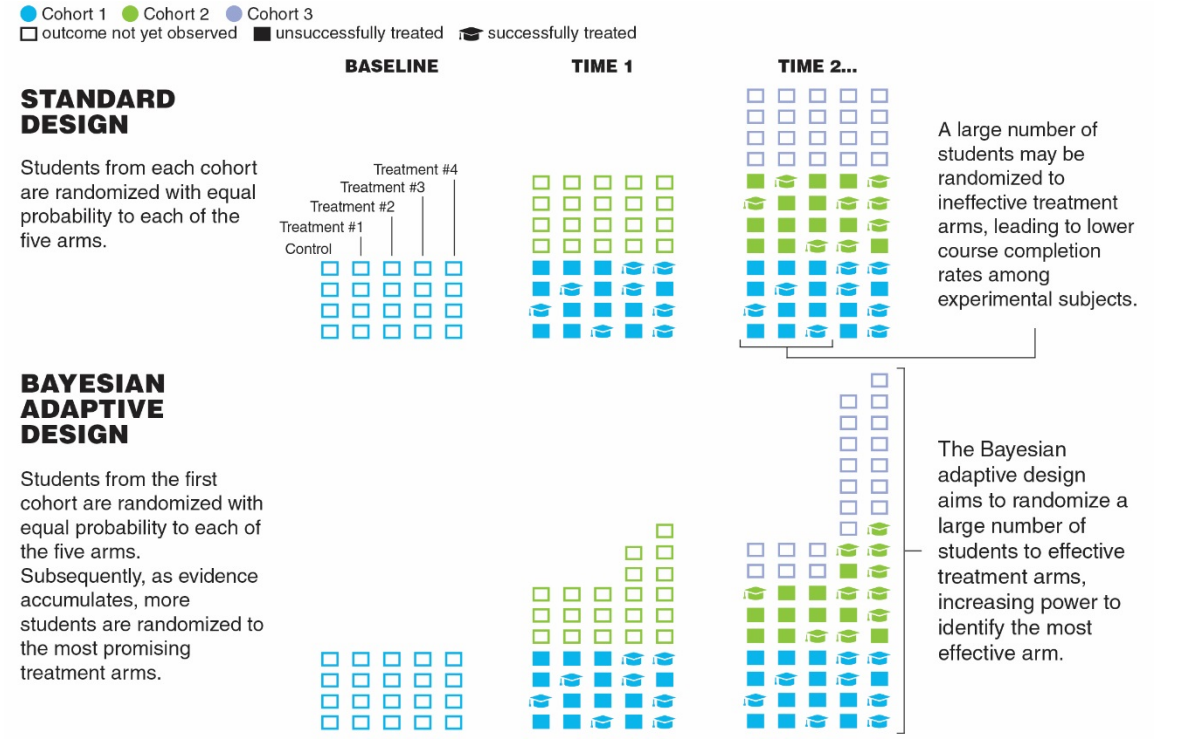
Study design

Using data and impact findings from a recent study of an informational nudge designed to increase course-completion in a massive online open course (MOOC), we simulate a series of experiments to examine the benefits and limitations of Bayesian adaptive design compared to the standard approach of randomizing with equal probability to each treatment arm throughout a study. Martinez (2015) conducted a nonadaptive randomized trial of students in a massive open online course to test whether changes in the way programs communicate with students can improve course completion rates. The RCT generated vast amounts of data on more than 23,000 course participants from 169 countries. In the final week of the course, students in the intervention arm received an email 'nudge' describing the negative correlation between procrastination and achievement. These students were 17 percent more likely to complete the course successfully the following week than students in the control group ($p < 0.01$). Additionally, in a post-hoc analysis, Martinez found that the treatment effect was heterogeneous across countries. For example, Germans assigned to the intervention arm were 167 percent more likely to complete the course ($p = 0.04$), but no effect was found for students from the United States ($p = 0.94$).

Building on Martinez's findings, this simulation study assumes four potential treatments that have different effects in different countries. To assess the performance of the Bayesian adaptive design relative to a traditional design, we compare two possible five-arm designs. The first design is the traditional design where students in each cohort are assigned with 20% probability to each of the five study arms (four treatments and one control). The second design is an adaptive design where the assignment probabilities for later cohorts vary based on initial findings. Each student will have an increased probability of being assigned to the treatment that is most effective in his or her country. To assess the performance of these designs under different conditions, we simulate 9 different scenarios, combining small, medium and large differences in

treatment effects between effective and ineffective treatment arms and small, medium, and large sample sizes. In each scenario, we simulate and analyze 1,000 synthetic data sets using the standard approach and 1,000 synthetic data sets using the proposed approach.

In this particular MOOC, new cohorts begin every two weeks and the course lasts for 6 weeks. This means that course completion outcomes for each cohort are observed in week four of the subsequent cohort, before the nudge is delivered in week 5. As the figure below shows, under the standard design, each cohort is assigned with 20% probability to each of the five study arms. Under the adaptive design, these probabilities change as the results from the previous cohort are observed. Treatments 4 and 5 are more successful at promoting completion, so more sample members are assigned to these arms and fewer to the less effective treatments or control.



Results

The Bayesian adaptive design successfully assigns more students to more effective treatment arms during the trial. In all nine scenarios, participants in the study are more likely to receive a beneficial treatment, increasing average outcomes overall for the set of study participants. This may help reduce concerns that rigorous studies are unfair because they withhold beneficial programs.

The Bayesian adaptive design produces better final inference than the standard design – students who enroll after the study concludes would have better outcomes, on average, under the Bayesian design. We quantify the quality of the final inference using a measure of predictive performance. Inference from the Bayesian adaptive design outperforms the standard design by this metric in all nine scenarios. These gains are achieved because the Bayesian adaptive design—by concentrating sample size in those treatment arms that seem most promising—achieves more power to compare successful treatments that differ in effectiveness

only slightly. This higher power enables the Bayesian adaptive design to distinguish among successful treatment arms, ultimately identifying the most effective.

The Bayesian design learns earlier and with smaller sample sizes what works for whom. Using the same metric of predictive performance, we compare inference produced by the Bayesian adaptive design after each cohort of students is enrolled to the inference produced by the standard design after the full experiment. Inference from the Bayesian design does as well (or better) with 6-7 cohorts of students as the standard design after enrolling all 26 cohorts. This means that studies using the Bayesian design could be completed much more quickly and with only about one-quarter of the sample, reducing study costs and timelines.

References

Barker, A.D., Sigman, C.C., Kelloff, G.J., Hylton, N.M., Berry, D.A., & Esserman, L.J. (2009). I-SPY 2: An adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. Clinical Pharmacology & Therapeutics, 86, 97–100.

Kim, E.S., Herbst, R.S., Wistuba, I.I., Lee, J.J., Blumenschein, G.R., Tsao, A., Hong, W.K. (2011). The BATTLE Trial: Personalizing therapy for lung cancer. Cancer Discovery, 1, 44–53.

Martinez, I. (2015). Never put off till tomorrow? Manuscript in preparation.