A Bayesian Perspective on Methodologies for Drawing Causal Inferences in Experimental and Non-Experimental Settings

David Kaplan
Cassie J. S. Chen
Department of Educational Psychology
THE UNIVERSITY of WISCONSIN MADISON

February 28, 2010

- This research was supported by a Vilas Associates Award from the Graduate School of the University of Wisconsin - Madison

Introduction

- Over the last decade significant advances have been made in the area of Bayesian statistical inference, owing mostly to computational developments and readily available software (see e.g. Gilks, Richardson, & Spiegelhalter, 1996).
- With these advances have come important applications of Bayesian methods to problems in the social and behavioral sciences.
- However, few applications of Bayesian methods to educational problems can be found, with the exception of Bayesian statistical methods for item response theory models in educational measurement (e.g. Fox & Glas, 2001).

- The importance of examining statistical modeling in the educational sciences from a Bayesian perspective cannot be overestimated.
- For too long, statistical methods applied to educational problems have rested on frequentist statistical hypothesis testing, originally developed by Fisher (1941/1925), and then later by Neyman & Pearson (1928).
- These approaches have been criticized as logically incoherent and that the Neyman-Pearson approach to hypothesis testing in particular has possibly done considerable damage to progress in the social and behavioral sciences. For interesting discussions on this problem, see Harlow, Mulaik, and Steiger (1994).
- An internally consistent and coherent alternative to the Neyman-Pearson paradigm in statistics lies with the Bayesian school. The Bayesian alternative to statistical inference provides a rational approach to incorporating uncertainty in statistical models.

- The frequentist perspective holds that parameters are fixed, and estimates of these fixed parameters are obtained from sample data.

- The Bayesian perspective, on the other hand, holds that because parameters are unknown, they are subject to the laws of probability, and sensible probability models can be formed to describe their behavior (see e.g. Box and Tiao,1973; and Gelman, Carlin, Stern, and Rubin (2003).

- Also, the frequentist perspective does not acknowledge that models themselves are sampled from a larger universe of possible models, none of which are true in any sense of the word.

- The Bayesian approach, on the other hand, attempts to determine which model is favored by the data (Hoeting, Madigan, Raftery, & Volinsky, 1999; Kass and Raftery, 1995).

- This talk will start with some preliminary “hot off the press” results from a Bayesian propensity score analysis.
- Then we’ll talk about possible extensions to Bayesian ANOVA.
- We’ll close with a discussion of advantages and disadvantages to the Bayesian approach.

- The main focus of BPSA is the recognition that the parameters of propensity score model are unknown and random and hence a Bayesian propensity score equation can be formed and combined with a Bayesian or frequentist causal model.
- Work on BPSA includes McCandless, Gustafson, and Austin (2009) who estimated the a Bayesian propensity score equation and causal model simultaneously.
- This has been criticized (rightly, we believe) for resulting in the treatment informing the estimation of the parameters of the causal equation.
- Violates the spirit of the propensity score.

Real data example of BPSA

- Data come from the ECLS-K. The “treatment” is full versus part day kindergarten attendance on IRT reading scores for children at the end of kindergarten.

- We randomly sampled 2000 children proportional to the number of children in full or part day kindergarten.

- This resulted in 1074 children in full day kindergarten and 926 children in part day kindergarten.

- Thirteen covariates were chosen for the propensity score equation. These included

1. Gender,
2. Race,
3. Child's learning style,
4. Self-control,
5. Social interactions,
6. Sadness/lonliness,
7. Impulsivness/overreactiveness,
8. Mother's employment status,
9. First time kindergartner, in 1998,
10. Mom's employment between birth and kindergarten,
11. Non-parental care arrangements,
12. SES ,
13. Number of grandparents who live close by.

- First we compare the unadjusted treatment effect using OLS compared to a Bayesian regression via MCMCregress.
- Next, we estimate the propensity score equation via a Bayesian generalized linear model with vague priors using MCMClogit.
- Then, we compare conventional PSA approaches (stratification, weighting and optimal matching) to the same approaches but with Bayesian estimated propensity scores. OLS is used for the causal model.
- Missing data was handled using multiple imputation with chained equations via the R program “MICE”

- For this talk, we use the R package "MCMClogit" to simulate from the posterior distribution of a logistic regression using a random walk Metropolis algorithm.

- The model is

$$y_i \sim Bern(\pi_i) \tag{1}$$

where via the inverse-link function

$$\pi_i = \frac{exp(\mathbf{z}_i'\boldsymbol{\beta})}{1 + exp(\mathbf{z}_i'\boldsymbol{\beta})}. \tag{2}$$

- An improper uniform prior for $\boldsymbol{\beta}$ was used reflecting a lack of information about the parameters.

- Note that conjugate priors with lesser or greater degrees of precision can be used for the propensity score model.

- For the Bayesian regression, “MCMCregress” simulates from the posterior distribution of the causal model using Gibbs sampling. The model is

$$y_i = x_i'\gamma + \epsilon_i, \tag{3}$$

where x_i is the treatment indicator and γ is the causal effect. In addition, we assume

$$\epsilon_i \sim N(0, \sigma^2) \tag{4}$$

- A uniform prior was used for the causal effect γ and an inverse gamma prior was used for σ^2, with shape parameter and scale parameter both 0.001.
- These are both non-informative priors

Propensity Score Methods

- We compare three conventional approaches for propensity score adjustments to their counterparts using Bayesian methods.
- For the conventional approaches we compute
  1. Strata sub-classification on the propensity score (standard approach using the R program *GLM*)
  2. Propensity score weighting (a special R function was written for this approach)
  3. Optimal matching (using the R progam *optmatch*)
- The Bayesian approaches simply utilize propensity scores obtained from the Bayesian logit models.
- MCMC sampling used 5000 burnin and 100,000 iterations with thinning interval of 10.
- Trace plots showed good convergence.

Results of PSA versus BPSA

Table 1
Final Result of PSA and BPSA

		Simple Linear Regression	**Bayesian[1] Regression**
Without Adjustment	Trt effect (SE)	0.589 (0.465)	0.592 (0.467)
	Confidence /Credible Interval	(-0.701, 1.880)	(-0.703, 1.887)
		PSA	**BPSA[2]**
Stratification	Trt effect (SE)	0.811 (0.472)	0.786 (0.497)
	Confidence /Credible Interval	(-0.496, 2.119)	(-0.591, 2.163)
Weighting	Trt effect (SE)	0.971 (0.462)	0.912 (0.510)
	Confidence /Credible Interval	(-0.311, 2.252)	(-0.501, 2.325)
Optimal Matching	Trt effect (SE)	0.897 (0.498)	0.892 (0.571)
	Confidence /Credible Interval	(-0.481, 2.275)	(-0.689, 2.473)

[1] Based on Bayesian regression assuming an improper uniform prior on the beta vector, and an inverse Gamma prior on the conditional error variance.

[2] Bayesian propensity score model using OLS regression for causal model.

Note: The final result of regression coefficient is just the mean of regression coefficients across five imputed data sets, while the final results of standard error and confidence/credible interval are calculated by Rubin's 1987 formula, taking the imputation variation into account.

Conclusions from BPSA

- As expected, without adjustment and with non-informative priors, the Bayesian treatment effect and credibility interval are virtually identical to the OLS results.
- The frequentist PSA versus Bayesian PSA provide similar treatment effects but the latter yield wider credibility intervals. A similar result was found in McCandliss, Gustafson, and Austin (2009).
- We note however, that the notion of statistical significance is not the key focus of the Bayesian approach.
- The credibility intervals indicate where the true treatment effect lies. This differs from the frequentist interpretation of the confidence interval.
- Current work is examining the effects of different choices of priors on the treatment effects and the width of the credibility intervals.

Inequality Constrained Bayesian ANOVA

- Can we bring Bayesian thinking into the experimental design framework?
- It has been argued everyone is a Bayesian at the design stage of an experiment. But what about the analysis stage?
- Over the last several years, the problem of direct model testing in the context of ANOVA methods for experimental designs has been discussed in the Bayesian literature under the concept of "inequality constrained ANOVA" (Hoijtink, Klugkist, & Boelen, 2008)

- The goal is not null hypothesis testing per se, but rather the estimation of models that incorporate "informative hypotheses" and the selection of a model that is favored by the data in hand.

- Following Klugkist (2008),

- $H_{1a}\,(\mu_1, \mu_2, \mu_3)$ is referred to as the "encompassing model" because it encompasses all of the constrained hypotheses.

- $H_{2a}\,(\mu_1 \approx \mu_2 \approx \mu_3)$ represents the Bayesian approach to frequentist null hypothesis testing, and can incorporate Cohen's "nil hypothesis" $|\mu_i - \mu_j| < \delta_{ij}$

- $H_{3a}\,(\mu_1 > \mu_2 > \mu_3)$ represents an ordered inequality constrained hypothesis presumably guided by theory.

Estimation Model Parameters

- The goal of Bayesian statistical inference is to obtain a complete description of the posterior distribution of all the parameters of the model. The goal of Bayesian ANOVA with informative hypotheses is no different.
- Each hypothesis under consideration generates its own posterior distribution of model parameters.
- These posterior distributions are sampled via MCMC methods and a complete summary is formed including posterior probability intervals.
- Model selection measures are then used to choose among the competing models that which is most closely aligned with the data.

Model Selection

- In conventional NHST, a *p*-value is obtained that provides the probability that the obtained effect is greater than or equal to *prespecified* α under the assumption that the null hypothesis is true.
- This approach to hypothesis testing is in stark contrast to the Bayesian perspective, which focuses generally on selecting the best model from a set of theoretically specified models after being confronted with the data.
- Specifically, Bayesian model selection examines the posterior probability for support in the data for the hypotheses of interest.

- The method of model selection in the context of Bayesian ANOVA with informative hypotheses utilizes Bayes factors. The Bayes factor provides a way to quantify the odds that the data favor one hypothesis over another. The Bayes factor can be written as

$$\frac{p(M_1|Y)}{p(M_2|Y)} = BF \times \left[\frac{p(M_1)}{p(M_2)}\right] \tag{5}$$

where the Bayes factor (BF) is defined as

$$BF = \frac{p(Y|M_1)}{p(Y|M_2)} = \frac{\int p(\theta_1|M_1)p(Y|\theta_1, M_1)d\theta_1}{\int p(\theta_2|M_2)p(Y|\theta_2, M_2)d\theta_2} \tag{6}$$

- The left-hand-side of equation (5) is the posterior probability of the data favoring M_1 over M_2.

- This posterior probability is related to the prior odds $p(M_1)/p(M_2)$ of the data favoring M_1 over M_2 weighted by the marginal likelihoods $p(Y|M_1)/p(Y|M_2)$ obtained in equation (6).

- Rearranging the terms in equation (5) allows us to express the Bayes factor in a more standard form. Specifically,

$$BF = \frac{p(M_1|Y)/p(M_1)}{p(M_2|Y)/p(M_2)} \tag{7}$$

Assuming neutral prior odds, i.e. $p(M_1) = p(M_2) = 1/2$, then the Bayes factor is simply the likelihood ratio.

- Rules of thumb have been developed to assess the quality of the evidence favoring one hypothesis over another using Bayes factors. Denoting M_1 as the reference model,

  - $BF \geq 1$: M_1 is supported.
  - $1 > BF \geq 10^{-\frac{1}{2}}$: minimal evidence against M_1.
  - $10^{-\frac{1}{2}} > BF \geq 10^{-1}$: substantial evidence against M_1.
  - $10^{-1} > BF \geq 10^{-2}$: strong evidence against M_1.
  - $10^{-2} > BF$: decisive evidence against M_1.

General Discussion

- Bayesian theory is not new. In fact, it predates frequentist theory by about 150 years.
- However, Bayesian computation is very new, in comparison to its frequentist counterpart.
- Extensions of frequentist methods (ANOVA, regression, SEM, HLM, etc.) to the Bayesian approach are happening at a rapid pace.
- What is unclear is whether we are seeing an interest in just another new estimator, or whether there is a genuine interest in the epistemological differences between the approaches and the practical consequences of those differences.

- We argue that the focus should be on the epistemological differences and a thorough examination of their practical consequences for educational research.

- In the context of causal inference in education (or elsewhere), the Bayesian approach offers a different perspective worth examining closely:

  1. Parameter and model uncertainty is explicitly handled.
  2. Priors on parameters and models can be elicited via meta analyses and expert opinion.
  3. Models can be selected on the basis of their predictive quality, not on statistical significance, per se.
  4. The true causal effect can be located in a range. The notion of hypothetical replications of an experiment, characteristic of frequentist statistics, is not needed.

- These advantages must be weighted against disadvantage.
  1. Bayesian inference is hard insofar as elicitation of priors is difficult.
  2. MCMC sampling can take a very long time!!
     - In our study, BPSA for stratification and weighting did not take very long.
     - Optimal matching took about 20 hours for one imputed data set, because each of the 2000 children had 10k runs that produce 10k propensity scores.
     - We expect that analysis time would be quite a bit shorter for more realistic quasi-experimental studies.
- Much more basic and applied research on Bayesian methods in educational settings is clearly needed to fully realize its potential.