The Bayesian Revolution and its Implications for Research With International Large-Scale Assessments

David Kaplan

Department of Educational Psychology

THE UNIVERSITY of WISCONSIN MADISON

Keynote Address: IEA IRC 2015, Cape Town, South Africa

- I would like to thank

- Leslie Rutkowski
- Paulina Koršňáková
- Sarah Howie

The Reverend Thomas Bayes, 1701–1761

Pierre-Simon Laplace, 1749–1827

> *[I]t is clear that it is not possible to think about learning from experience and acting on it without coming to terms with Bayes' theorem.* — *Jerome Cornfield*

Introduction

- Bayesian statistics has long been overlooked in the quantitative methods training of social scientists.

    1. Until recently, it was not feasible to conduct statistical modeling from a Bayesian perspective owing to its complexity and lack of availability.

    2. Bayesian statistics represents a powerful alternative to frequentist (classical) statistics, and is therefore, controversial.

- There is now a renaissance in the development and application of Bayesian statistical methods due to software developments.

- What role can Bayesian inference play in improving research using large-scale assessments?

Outline

1. Overview paradigm differences between frequentist and Bayesian statistics.
2. Introduce the central concepts of Bayesian statistics.
3. Discuss Bayesian hypothesis testing and model buidling.
4. Provide an example using data from PIRLS.
5. Implications for IEA large-scale assessments.
6. Final thoughts

Paradigm Differences

- For frequentists, the basic idea is that probability is represented by the model of **long run frequency**.
- Frequentist probability underlies the Fisher and Neyman-Pearson schools of statistics – the conventional methods of statistics we most often use.
- The frequentist formulation rests on the idea of equally probable and stochastically independent events
- The physical representation is the coin toss, which relates to the idea of a very large (actually infinite) number of repeated experiments.

- The entire structure of Neyman - Pearson hypothesis testing and Fisherian statistics (together referred to as the **frequentist school**) is based on frequentist probability.

- Our conclusions regarding null and alternative hypotheses presuppose the idea that we could conduct the same experiment an infinite number of times.

- Our interpretation of confidence intervals also assumes a fixed parameter and CIs that vary over an infinitely large number of identical experiments.

- But there is another view of probability as **subjective belief**.
- The physical model in this case is that of the “bet”.
- Here, probability is not based on an infinite number of repeatable and stochastically independent events, but rather on how much knowledge you have and how much you are willing to bet.
- Subjective probability allows one to address questions such as “what is the probability that my team will win the World Cup?” Relative frequency supplies information, but it is not the same as probability and can be quite different.
- This notion of subjective probability underlies Bayesian statistics.

Bayes' Theorem

- Consider the joint probability of two events, Y and X, for example observing lung cancer and smoking jointly.
- The joint probability can be written as

$$p(cancer, smoking) = p(cancer|smoking)p(smoking) \tag{1}$$

- Similarly

$$p(smoking, cancer) = p(smoking|cancer)p(cancer) \tag{2}$$

- Because these are symmetric, we can set them equal to each other to obtain the following

$$p(cancer|smoking)p(smoking) = p(smoking|cancer)p(cancer) \tag{3}$$

$$p(cancer|smoking) = \frac{p(smoking|cancer)p(cancer)}{p(smoking)} \tag{4}$$

- The inverse probability theorem (Bayes' theorem) states

$$p(smoking|cancer) = \frac{p(cancer|smoking)p(smoking)}{p(cancer)} \tag{5}$$

Statistical Elements of Bayes' Theorem

- What is the role of Bayes' theorem for statistical inference?
- Denote by Y a random variable that takes on a realized value y such as a student's reading proficiency.
- Denote by θ a parameter that we believe characterizes the probability model for y.
- We are concerned with determining the probability of observing y given the unknown parameters θ, which we write as

$$p(y|\theta). \tag{6}$$

- The key difference between Bayesian statistical inference and frequentist statistical inference concerns the nature of the unknown parameters θ.

- In the frequentist tradition, the assumption is that θ is unknown and has a fixed value that we wish to estimate. Our uncertainty about θ is not taken into account in the frequentist tradition.

- In Bayesian statistical inference, θ is also considered unknown so we specify a probability distribution that reflects our uncertainty about the true value of θ.

- Because both the observed data y and the parameters θ are assumed to be random variables, we can model the joint probability of the parameters and the data as a function of the conditional distribution of the data given the parameters, and the prior distribution of the parameters.

- More formally,

$$p(\theta, y) = p(y|\theta)p(\theta). \tag{7}$$

where $p(\theta, y)$ is the joint distribution of the parameters and the data. Following Bayes' theorem described earlier, we obtain

Bayes' Theorem

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}, \tag{8}$$

where $p(\theta|y)$ is referred to as the *posterior distribution* of the parameters θ given the observed data y.

- Sometimes we see Bayes' theorem written as

$$p(\theta|y) \propto p(y|\theta)p(\theta) \tag{9}$$

- The prior represents what is reasonable to believe about a parameter prior to data collection.
- But how do we choose a prior?
  - Leamer (1983) orders priors on the basis of degree of confidence. Leamer's hierarchy of confidence is as follow: truths (e.g. axioms) $>$ facts (data) $>$ opinions (e.g. expert judgement) $>$ conventions (e.g. pre-set alpha levels).
  - Research with LSAs will rely more or less on data and expert opinion.
- Moderation of our prior beliefs by the data in hand is the key meaning behind equations (8) and (9).

Non-informative priors

- In some cases we may not be in possession of enough prior information to aid in drawing posterior inferences.
- From a Bayesian perspective, this lack of information is still important to consider and incorporate into our statistical specifications.
- In other words, it is equally important to quantify our ignorance as it is to quantify our cumulative understanding of a problem at hand.
- The standard approach to quantifying our ignorance is to incorporate a non-informative prior into our specification.
- The uniform prior over some range is a routinely used non-informative prior.

Informative Priors

- It may be the case that some information can be brought to bear on a problem and be systematically incorporated into the prior distribution.
- Such "subjective" priors are called *informative*.
- Subjective priors be elicited from highly "objective" sources or highly "subjective" sources.
- Both types of priors are brought into a Bayesian analysis by specifying the prior distribution $p(\theta)$ and its parameters (so-called *hyperparameters*).

Bayesian Hypothesis Testing

- LSA research naturally focuses on hypothesis testing.
- The approach most widely used in the social and behavioral sciences is the Neyman-Pearson approach.
- An interesting aspect of the Neyman-Pearson approach to hypothesis testing is that students (as well as many seasoned researchers) appear to have a very difficult time grasping its principles.
- Gigerenzer (2004) argued that much of the difficulty in grasping frequentist hypothesis testing lies in the conflation of Fisherian hypothesis testing and the Neyman-Pearson approach to hypothesis testing.

- Fisher's early approach to hypothesis testing required specifying only the null hypothesis.
- A conventional significance level is chosen (usually the 5% level).
- Once the test is conducted, the result is either significant ($p < .05$) or it is not ($p > .05$).
- If the resulting test is significant, then the null hypothesis is rejected. However, if the resulting test is not significant, then no conclusion can be drawn.
- Fisher's approach was based on looking at how data could inform evidence for a hypothesis.

- The Neyman and Pearson approach requires that two hypotheses be specified – the null and alternative hypothesis – and is designed to inform specific sets of actions. It's about making a choice, not about evidence against a hypothesis.
- By specifying two hypotheses, one can compute a desired tradeoff between two types of errors: Type I errors (α) and Type II errors (β)
- The goal is not to assess the evidence against a hypothesis (or model) taken as true. Rather, it is whether the data provide support for taking one of two competing sets of actions.

- The conflation of Fisherian and Neyman-Pearson hypothesis testing lies in the use and interpretation of the p-*value*.
- In Fisher's paradigm, the p-value is a matter of convention with the resulting outcome being based on the data.
- In the Neyman-Pearson paradigm, α and β are determined prior to the experiment being conducted and refer to a consideration of the cost of making one or the other error.
- Indeed, in the Neyman-Pearson approach, the problem is one of finding a balance between α, power, and sample size.
- The Neyman-Pearson approach is best suited for experimental planning where the interest would be in error control (Type I or Type II).

- However, this balance is virtually always ignored and α = 0.05 is used.
- The point is that the p-value and α are not the same thing.
- This can lead a researcher to set α ahead of time, as per the Neyman-Pearson school, but then communicate a different level of "significance" after running the test.
- The conventional practice is even worse than described, as evidenced by nonsensical phrases such as results "trending toward significance".

- So what is the p-value ... really?

$$p - value = p(D|H_0) \tag{10}$$

- To compute this probability, we calculate a test statistic, $T(x)$
- We compare $T(x)$ obtained from the data to a hypothetical distribution of T based on a hypothetically infinite number of identically conducted experiments (or studies).
- This allows us to define a rejection region.
- The p-value is the probability of obtaining a test statistic as extreme or more extreme than what actually observed under the assumption that the null hypothesis (which is never true) is true.

- In the Bayesian framework the concern is

$$p(H_0|D) \tag{11}$$

- The Bayesian goal is to see if the data in hand provide evidence for a hypothesis of interest.
- Typically, H_0 is described by parameters, so this is the posterior distribution.
- Given priors, the Bayesian approach allows us to assess the strength of evidence for a hypothesis taking into account uncertainty in the parameters.

Bayesian Model Building

- The frequentist and Bayesian goals of model building are the same.

  1. Specify an initial model relying on a lesser or greater degree of prior theoretical knowledge.
  2. Fit the model to data.
  3. Evaluate the quality of the model (and possibly respecify).
  4. Choose the “best model” will be chosen for some purpose.

- Despite these similarities there are important differences.
  1. Specifying a full probability model for the data and priors
  2. Choosing models from a predictive point of view.

Bayes Factors

- A very simple and intuitive approach to model building and model selection uses so-called *Bayes factors* (Kass & Raftery, 1995)
- In essence, the Bayes factor provides a way to quantify the odds that the data favor one hypothesis over another. A key benefit of Bayes factors is that models do not have to be nested.
- We can think of regression models with a different number of variables, or two structural equation models specifying very different directions of mediating effects.

- The goal is to develop a quantity that expresses the extent to which the data support one model over another.
- We weight the odds ratio of the data given the model by our prior belief that one model may be superior to another.
- This weighting gives rise to our updated view of evidence provided by the data for either hypothesis.
- The updated (or posterior) odds is the Bayes factor.
- In practice, there may be no prior preference for one model over the other. In this case, we use neutral prior odds.

Bayesian Model Averaging

- The distinctive feature that separates Bayesian statistical inference from its frequentist counterpart is its focus on describing and modeling all forms of uncertainty.
- The primary focus of uncertainty within a Bayesian modeling exercise concerns prior knowledge about model parameters.
- In the Bayesian framework, all unknowns are described by probability distributions.
- Parameters constitute the central focus of statistical modeling, and because they are unknown, Bayesian inference encodes background knowledge about parameters in the form of the prior distribution.

- Within the Bayesian framework, parameters are not the only unknown elements.
- The Bayesian framework recognizes that models themselves possess uncertainty insofar as a particular model is typically chosen based on prior knowledge of the problem at hand, and the variables that have been used in previously specified models.
- This form of uncertainty often goes unnoticed.

- Quoting Hoeting et al. (1999)

  > *"Standard statistical practice ignores model uncertainty. Data analysts typically select a model from some class of models and then proceed as if the selected model had generated the data. This approach ignores the uncertainty in model selection, leading to over-confident inferences and decisions that are more risky than one thinks they are."(pg. 382)*

- An internally consistent Bayesian framework for structural equation modeling and estimation must also account for model uncertainty.

- The current approach to addressing the problem of uncertainty lies in the method of *Bayesian model averaging* (BMA).

- The essential steps of BMA are as follows

  1. Specify a model of interest.
  2. Search the space of all possible models that satisfy certain criteria of parsimony.
  3. Estimate each sub-model and obtain each model's posterior probability.
  4. Create a new set of coefficients based on a weighted average of the coefficients across all sub-models weighted by their PMPs.
  5. Use the Bayesian model averaged coefficients for prediction.

- Prediction based on BMA will be better than the original model or any single sub-model.

Example: Bayesian Linear Regression

- Consider a regression of reading scores from n students on the PIRLS reading assessment explained by a set of demographic, attitudinal, and school environment characteristics.

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \tag{12}$$

  where $\boldsymbol{\beta}$ is a vector of regression coefficients (including the intercept).

- We assume that student level PIRLS reading scores scores are generated from a normal distribution.

- We also assume that the error terms are independently, identically, and normally distributed.

- Recall that the components of Bayes' theorem require the specification of the model and the priors on all model parameters.
- We will write the regression model as

$$p(\boldsymbol{X}, \boldsymbol{y} | \boldsymbol{\beta}, \sigma^2) \tag{13}$$

- Conventional statistics stops here and estimates the model parameters with either maximum likelihood estimation or ordinary least squares.
- But for Bayesian regression we need to specify the priors for all model parameters.

Non-informative Priors

- First consider non-informative uniform priors.
- For the regression coefficients and error variance, we will use the uniform prior.
- The joint posterior distribution of the model parameters written as

$$p(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}, \boldsymbol{X}) \propto p(\boldsymbol{X}, \boldsymbol{y} | \boldsymbol{\beta}, \sigma^2) \times p(\boldsymbol{\beta}) \times p(\sigma^2). \quad (14)$$

- What about informative priors?
- These can be elicited from subjective judgement, expert opinion, or previous research
- We usually use the normal distribution for the prior on the regression coefficients centered at what we think the mean should be along with a measure of our precision.
- For the variance of the disturbance term we will typically use a non-informative prior.
- From here, we can obtain the joint posterior distribution of all model parameters.

MCMC

- The key reason for the increased popularity of Bayesian methods in the social and behavioral sciences has been the (re)-discovery of numerical algorithms for estimating the posterior distribution of the model parameters given the data.

- Prior to these developments, it was virtually impossible to analytically derive summary measures of the posterior distribution, particularly for complex models with many parameters.

- Rather than attempting the impossible task of analytically solving for estimates of a complex posterior distribution, we can instead draw samples from the posterior distribution directly and summarize the distribution formed by those samples.
- This is referred to as *Monte Carlo integration*.
- This methods is called *Markov Chain Monte Carlo* sampling and is now widely available in software such as Mplus, Stata, OpenBUGS, JAGS, and numerous functions in R

A Cautionary Note About MCMC

- MCMC requires a number of decisions related to the computations.
- MCMC can take a long time to run if the models are complicated.
- It is critically important to examine convergence diagnostics.
- Using MCMC as a method to solve analytic problems employs non-informative priors as defaults, and these priors can, in principle, have a big effect.

Table 1: Bayesian Linear Regression Estimates: Non-informative Prior Case

Parameter	EAP	SD	95% PPI
Full Model			
INTERCEPT	346.14	6.96	332.57, 359.59
READING on male	0.02	1.59	-3.02, 3.18
READING on ASBG04	9.91	0.75	8.42, 11.39
READING on ASBGSBS	4.30	0.42	3.49, 5.13
READING on ASBGSLR	7.66	1.32	6.79, 8.57
READING on ASBGSMR	-5.91	0.51	-6.90, -4.90
READING on ASBGSCR	11.23	0.44	10.36, 12.08
READING on ASBR05E	5.99	1.34	3.41, 8.64
READING on ASBR05F	-5.04	1.42	-7.86, -2.28
READING on ASBR05G	-3.05	1.32	-5.67, -0.36

Note. male (1=male); ASBG04=# of books in home; ASBGSBS=Bullied at school; ASBGSLR = Students like reading; ASBGSMR = Students motivated to read; ASBGSCR = Students confidence in their reading; ASBR05E = Teacher is easy to understand; ASBR05F = Interested in what teacher says; ASBR05G = Teacher gives interesting things to read.

Table 2: Bayesian Regression Estimates: Informative Priors based on Linear Regression

Parameter	EAP	SD	95% PPI
Full Model			
INTERCEPT	399.61	0.43	398.76, 400.44
READING on male	-0.03	0.44	-0.86, 0.84
READING on ASBG04	9.26	0.37	8.52, 10.00
READING on ASBGSBS	3.03	0.27	2.50, 3.55
READING on ASBGSLR	6.97	0.29	6.41, 7.55
READING on ASBGSMR	-7.11	0.31	-7.71, -6.52
READING on ASBGSCR	10.39	0.29	9.85, 10.96
READING on ASBR05E	6.08	0.41	5.29, 6.89
READING on ASBR05F	-6.46	0.42	-7.31, -5.64
READING on ASBR05G	-4.50	0.41	-5.29, -3.68

Note. male (1=male); ASBG04=# of books in home; ASBGSBS=Bullied at school; ASBGSLR = Students like reading; ASBGSMR = Students motivated to read; ASBGSCR = Students confidence in their reading; ASBR05E = Teacher is easy to understand; ASBR05F = Interested in what teacher says; ASBR05G = Teacher gives interesting things to read.

Table 3: Natural Log Bayes Factors: Informative Priors Case

Sub-Model	FULLMOD	BGRND	POSREAD	SCHENV
FULLMOD	0	3044	88	1088
BGRND	-3044	0	-2956	-1957
POSREAD	-88	2956	0	999
SCHENV	-1088	1957	-999	0

Note. BGRND=(male, ASBG04=# of books in home); POSREAD=(ASBGSLR = Students like reading; ASBGSMR = Students motivated to read; ASBGSCR = Students confidence in their reading); SCHENV=(ASBGSBS= Bullied at school; ASBR05E = Teacher is easy to understand; ASBR05F = Interested in what teacher says; ASBR05G = Teacher gives interesting things to read.) Very strong evidence in favor of FULLMOD over all other models.

Table 4: Selected models from BMA

Parameter	M_1	M_2	M_3
male			
ASBG04	●	●	●
ASBGSBS	●	●	●
ASBGSLR	●	●	●
ASBGSMR	●	●	●
ASBGSCR	●	●	●
ASBR05E		●	
ASBR05F	●	●	
ASBR05G			
BIC	-819.968	-816.408	-814.161
PMP	0.817	0.138	0.045

Note. PMP = posterior model probability.

Table 5: Comparison of the result of BMA to the result of the Bayesian Regression Model

	BMA			Bayesian Regression Model			
Parameter	mean$(\beta\|y)$	SD$(\beta\|y)$	p$(\beta > 0\|y)\%$	EAP	SD	95%	PPI
male	0.00	0.00	0.0	0.08	2.15	-4.10	4.24
ASBG04	9.25	1.06	100.0	9.23	1.06	7.15	11.27
ASBGSBS	4.14	0.57	100.0	4.08	0.57	2.98	5.19
ASBGSLR	7.67	0.63	100.0	7.69	0.64	6.44	8.94
ASBGSMR	-5.03	0.69	100.0	-4.97	0.68	-6.30	-3.63
ASBGSCR	10.59	0.59	100.0	10.47	0.60	9.31	11.6
ASBR05E	0.53	1.49	13.8	4.12	1.87	0.48	7.83
ASBR05F	6.03	2.13	95.5	-6.95	1.97	-10.84	-3.05
ASBR05G	0.00	0.00	0.0	-1.12	1.87	-4.68	2.55

Note. EAP = expected a posterior; SD = posterior standard deviation; PPI = posterior probability interval.

Table 6: Comparison of predictive performance

Method	90% Coverage	Log score
BMA	93.26	-.0698
Best model from BMA	88.68	-.1201
Bayesian Regression Model	88.65	-.1205

Note. Log score, log score of predictive coverage.

Implications for IEA LSAs

- Bayesian methods are already an important part in LSAs.
- The methodology of plausible values rests on Bayesian ideas.
- New developments in Bayesian methodology are particularly useful for research with LSAs.

- Approximate measurement invariance for cross-country comparisons
- Bayesian model averaging for structural equation models (Kaplan & Lee, under review)
- Bayesian approaches to small area estimation (related to missing data issues).
- Bayesian methods for complex sampling designs.

Final Thoughts

- Bayesian statistics represents a powerful alternative to frequentist (classical) statistics, and is therefore, controversial.
- The controversy lies in differing perspectives regarding the nature of probability, and the implications for statistical practice that arise from those perspectives.
- The frequentist framework views probability as synonymous with long-run frequency, and that the infinitely repeating coin-toss represents the canonical example of the frequentist view. frequency.
- In contrast, the Bayesian viewpoint regards probability as the quantification of the “subjective” experience of uncertainty.

Summarizing the Bayesian Advantage

- The major advantages of Bayesian statistical inference over frequentist statistical inference are

  1. Coherence
  2. Flexibility in handling complex data structures
  3. Inferences based on data actually observed
  4. Quantifying evidence
  5. Incorporating prior knowledge.

  - **WE KNOW THINGS!**

Subjective v. Objective Bayes

- Subjective Bayesian practice attempts to bring prior knowledge directly into an analysis. This prior knowledge represents the analysts (or others) degree-of-uncertainity.
- An analyst's degree-of-uncertainty is encoded directly into the specification of the prior distribution, and in particular on the degree of precision around the parameter of interest.
- The advantages include
  1. Subjective priors have proper statistical properties
  2. Priors can be based on factual prior knowledge
  3. Small sample sizes can be handled.

- The disadvantages to the use of subjective priors according to Press (2003) are

  1. It is not always easy to encode prior knowledge into probability distributions.
  2. Subjective priors are not always appropriate in public policy or clinical situations.

- In terms of advantages Press (2003) notes that

  1. Objective priors can be used as benchmarks against which choices of other priors can be compared. Sensitivity analysis.
  2. Objective priors reflect the view that little information is available about the process that generated the data
  3. An objective prior provides results equivalent to those based on a frequentist analysis
  4. Objective priors are sensible public policy priors.

Final Thoughts: A Call for Evidenced-based Subjective Bayes

- "Subjectivism" within the Bayesian framework runs the gamut from the elicitation of personal beliefs to making use of the best available historical data available to inform priors.
- I argue along the lines of Jaynes (1968) – namely that the requirements of science demand reference to "specific, factual data on which ... opinions are based" (pg. 228).
- This view is also consistent with Leamer's hierarchy of confidence on which priors should be ordered.
- We may refer to this view as an *evidence-based* form of subjective Bayes which acknowledges (1) the subjectivity that lies in the choice of historical data; (2) the encoding of historical data into hyperparameters of the prior distribution; and (3) the choice among competing models to be used to analyze the data.

- What if factual historical data are not available?

- Berger (2006) states that reference priors should be used “in scenarios in which a subjective analysis is not tenable”, although such scenarios are probably rare, especially when applied to large-scale educational assessments.

- The goal, nevertheless, is to shift the practice of Bayesian statistics away from the elicitation of personal opinion (expert or otherwise) which could, in principle, bias results toward a specific outcome, and instead move Bayesian practice toward the warranted use prior objective empirical data for the specification of priors.

- The specification of any prior should be explicitly warranted against observable, empirical data and available for critique by the relevant scholarly community.

Conclusions

- Bayesian statistical inference is, arguably, superior to the frequentist school as a means of creating and updating new knowledge in education.

.

[I]t is clear that it is not possible to think about learning from experience and acting on it without coming to terms with Bayes' theorem. *- Jerome Cornfield*

- An evidence-based focus that ties the specification of priors to objective empirical data should provide stronger warrants for conclusions drawn from a Bayesian analysis.

- Predictive criteria rather than goodness-of-fit should always be used as a means of testing and choosing among Bayesian models.

> *All models are wrong, but some are useful.*
> *- George Box.*

- More research is needed, but the full benefit of the Bayesian approach to research in education will be realized when it is more widely adopted and yields reliable predictions that advance knowledge.

THANK YOU
DANKIE
DANKE
MERCI
GRAZIE
GRACIAS