# A Brief Introduction to Bayesian Statistics

David Kaplan

Department of Educational Psychology

THE UNIVERSITY of WISCONSIN MADISON

Bayesian Methods for Social Policy Research and Evaluation, Washington, DC 2017

# 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*

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

- Typically, the only introduction that a student might have to Bayesian ideas is a brief overview of Bayes' theorem while studying probability in an introductory statistics class.

  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.

- Bayesian statistical methods are now more popular owing to the development of powerful statistical software tools that make the estimation of complex models feasible from a Bayesian perspective.

# Outline



- Major differences between the Bayesian and frequentist paradigms of statistics
- Bayes' Theorem
- Priors
- Computation
- Bayesian hypothesis testing
- Bayesian model building and evaluation
- Debates within the Bayesian school.

# Paradigm Difference I: Conceptions of Probability



- 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 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".
- Consider the situation of betting on who will win the World Series.
- Here, probability is not based on an infinite number of repeatable and 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 Series?" 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

$$p(cancer|smoking)p(smoking) = p(smoking|cancer)p(cancer) \quad (3)$$

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

- The inverse probability theorem (Bayes' theorem) states

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

- Why do we care?
- Because this is how you could go from the probability of having cancer given that the patient smokes, to the probability that the patient smokes given that he/she has cancer.
- We simply need the marginal probability of smoking and the marginal probability of cancer.
- We will refer to these marginal probabilities as *prior probabilities*.

# 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 $y$.
- For example, a person's socio-economic status could be considered a random variable taking on a very large set of possible values.
- Once the person identifies his/her socioeconomic status, the random variable $Y$ is now realized as $y$.
- Because $Y$ is unobserved and random, we need to specify a probability model to explain how we obtained the actual data values $y$.

- Next, denote by $\theta$ a parameter that we believe characterizes the probability model of interest.
- The parameter $\theta$ can be a single number, such as the mean or the variance of a distribution, or it can be a set of numbers, such as a set of regression coefficients in regression analysis or factor loadings in factor analysis.
- We are concerned with determining the probability of observing $y$ given the unknown parameters $\theta$, which we write as $p(y|\theta)$.
- In statistical inference, the goal is to obtain estimates of the unknown parameters given the data.

# Paradigm difference II: Nature of parameters



- The key difference between Bayesian statistical inference and frequentist statistical inference concerns the nature of the unknown parameters $\theta$.
- In the frequentist tradition, the assumption is that $\theta$ is unknown, but that estimation of this parameter does not incorporate the our uncertainty about what is reasonable to believe about the parameter.
- In Bayesian statistical inference, $\theta$ is considered unknown and random, possessing a probability distribution that reflects our uncertainty about the true value of $\theta$.
- Because both the observed data $y$ and the parameters $\theta$ are assumed random, 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,

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

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

## Bayes' Theorem

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

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

- Equation (7) represents the core of Bayesian statistical inference and is what separates Bayesian statistics from frequentist statistics.

- Equation (7) states that our uncertainty regarding the parameters of our model, as expressed by the prior distribution $p(\theta)$, is *weighted* by the actual data $p(y|\theta)$ yielding an updated estimate of our uncertainty, as expressed in the posterior distribution $p(\theta|y)$.

# Priors



- Bayesian statistics requires that prior distributions be specified for **ALL** model parameters.
- There are three classes of prior distributions:
  1. Non-informative priors
     - We use these to quantify actual or feigned ignorance. Lets the data maximally speak.
  2. Weakly informative priors
     - Allows some weak bounds to be placed on parameters.
  3. Informative priors
     - Allows cumulative information to be considered.

# MCMC



- 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.

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

- Instead of analytically solving for estimates of a complex posterior distribution, we can instead draw samples from $p(\theta|y)$ and summarize the distribution formed by those samples. This is referred to as *Monte Carlo integration*.

- The most popular methods of MCMC are the Gibbs sampler, the Metropolis-Hastings algorithm, and more recently, the Hamiltonian MCMC algorithm.

# Bayesian Hypothesis Testing



- Bayesian hypothesis testing begins first by obtaining summaries of relevant distributions.
- The difference between Bayesian and frequentist statistics is that with Bayesian statistics we wish to obtain summaries of the posterior distribution.
- The expressions for the mean (EAP), variance, and mode (MAP) of the posterior distribution come from expressions for the mean and variance of conditional distributions generally.

# Posterior Probability Intervals



- In addition to point summary measures, it may also be desirable to provide interval summaries of the posterior distribution.
- Recall that the frequentist confidence interval requires that we imagine an infinite number of repeated samples from the population characterized by $\mu$.
- For any given sample, we can obtain the sample mean $\bar{x}$ and then form a $100(1 - \alpha)$% confidence interval.
- The correct frequentist interpretation is that $100(1 - \alpha)$% of the confidence intervals formed this usual way capture the true parameter $\mu$ under the null hypothesis. Notice that the probability that the parameter is in the interval is either zero or one.

# Posterior Probability Intervals (cont'd)



- In contrast, the Bayesian framework assumes that a parameter has a probability distribution.
- Sampling from the posterior distribution of the model parameters, we can obtain its quantiles. From the quantiles, we can directly obtain the probability that a parameter lies within a particular interval.
- So, a 95% posterior probability interval would mean that the probability that the parameter lies in the interval is 0.95.
- The posterior probability interval can, of course, include zero. Zero is a credible value for the parameter.
- Notice that this is entirely different from the frequentist interpretation, and arguably aligns with common sense.

# Bayesian Model Building and Evaluation



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

  1. Model specification based on prior knowledge
  2. Model estimation and fitting to data obtained from a relevant population
  3. Model evaluation and modification
  4. Model choice

- Despite these similarities there are important differences.
- A major difference between the Bayesian and frequentist goals of model building lie in the model specification stage.
- Because the Bayesian perspective explicitly incorporates uncertainty regarding model parameters using priors, the first phase of modeling building requires the specification of a full probability model for the data and the parameters of the model.
- Model fit implies that the full probability model fits the data. Lack of model fit may be due to incorrect specification of likelihood, the prior distribution, or both.

- Another difference between the Bayesian and frequentist goals of model building relates to the justification for choosing a particular model among a set of competing models.
- Model building and model choice in the frequentist domain is based primarily on choosing the model that best fits the data.
- This has certainly been the key motivation for model building, re-specification, and model choice in the context of structural equation modeling (Kaplan 2009).
- In the Bayesian domain, the choice among a set of competing models is based on which model provides the best posterior predictions.
- That is, the choice among a set of competing models should be based on which model will best predict what actually happened.

# Posterior Predictive Checking



- A very natural way of evaluating the quality of a model is to examine how well the model fits the actual data.
- In the context of Bayesian statistics, the approach to examining how well a model predicts the data is based on the notion of *posterior predictive checks*, and the accompanying *posterior predictive p-value*.
- The general idea behind posterior predictive checking is that there should be little, if any, discrepancy between data generated by the model (including the priors), and the actual data itself.
- Posterior predictive checking is a method for assessing the specification quality of the model. Any deviation between the data generated from the model and the actual data as measured by the posterior predictive distribution and the Bayesian $p$-value implies model misspecification.

# Bayes Factors



- A very simple and intuitive approach to model building and model selection uses so-called *Bayes factors* (Kass & Raftery, 1995)
- Consider two regression models with a different number of variables, or two structural equation models specifying very different directions of mediating effects.
- 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.
- The Bayes' factor gives rise to the popular Bayesian information criterion (BIC).

# Bayesian Model Averaging



- The selection of a particular model from a universe of possible models can also be characterized as a problem of uncertainty. This problem was succinctly stated by Hoeting, Raftery & Madigan (1999) who write

> *"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 approach to addressing the problem of model selection is the method of *Bayesian model averaging* (BMA).
- I will discuss this on Friday.

# Debates Within the Bayesian School



- 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.
- Recall that 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. .

# Summarizing the Bayesian Advantage



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

  1. Coherence – Probability axioms constrain how degrees-of-belief translate into action.
  2. Flexibility in handling very complex models (through MCMC)
  3. Inferences based on data actually observed
  4. Quantifying evidence
  5. Incorporating prior knowledge

# 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 prior distribution, and specifically in the degree of precision around the parameter of interest.
- The advantages according to Press (2003) include
  1. Subjective priors are proper
  2. Priors can be based on factual prior knowledge
  3. Small sample sizes can be handled. Though this is not a free lunch.

- 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.

- Within objective Bayesian statistics, focus is on letting the data maximally speak
- In terms of advantages of the objective Bayes school, Press (2003) notes that
    1. Objective priors can be used as benchmarks against which choices of other priors can be compared
    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.

- In terms of disadvantages of objective priors, Press (2003) notes that

  1. Objective priors can lead to improper results when the domain of the parameters lie on the real number line.
  2. Parameters with objective priors are often independent of one another, whereas in most multi-parameter statistical models, parameters are correlated.
  3. Expressing complete ignorance about a parameter via an objective prior leads to incorrect inferences about functions of the parameter.

- In its extreme form, the subjectivist school
  - allows for personal opinion to be elicited and incorporated into a Bayesian analysis.
  - places no restriction on the source, reliability, or validity of the elicited opinion as long as the rules of probability (coherence) are adhered to.
- In its extreme form, the objectivist school
  - views personal opinion as the realm of psychology with no place in a statistical analysis.
  - would require formal rules for choosing reference priors.

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

- The Bayesian school of statistical inference is:

  1. the only coherent approach to incorporating prior information thus leading to updated inferences and growth in knowledge.

  2. represents an internally consistent and rigorous way out of the $p$-value problem.

  3. approaches model building, evaluation, and choice, from a predictive point of view.

- As always, the full benefit of the Bayesian approach to research in the social sciences will be realized when it is more widely adopted and yields reliable predictions that advance knowledge.

# THANK YOU