Introduction to Bayesian Inference and Modeling

Edps 590BAY

Carolyn J. Anderson

Department of Educational Psychology
ILLINOIS

Fall 2019

Overview

- What is Bayes theorem
- Why Bayesian analysis
- What is probability?
- Basic Steps
- An little example
- History (not all of the 705+ people that influenced development of Bayesian approach)
- In class work with probabilities

Depending on the book that you select for this course, read either Gelman et al. Chapter 1 or Kruschke Chapters 1 & 2.

Main References for Course

Throughout the coures, I will take material from

- Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., & Rubin, D.B. (20114). *Bayesian Data Analysis*, 3rd Edition. Boco Raton, FL, CRC/Taylor & Francis.**
- Hoff, P.D., (2009). *A First Course in Bayesian Statistical Methods*. NY: Sringer.**
- McElreath, R.M. (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Boco Raton, FL, CRC/Taylor & Francis.
- Kruschke, J.K. (2015). *Doing Bayesian Data Analysis: A Tutorial with JAGS and Stan*. NY: Academic Press.**

** There are e-versions these of from the UofI library. There is a verson of McElreath, but I couldn't get if from UofI e-collection.

Bayes Theorem

A whole semester on this?

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

where

- y is data, sample from some population.
- θ is unknown parameter.
- $p(y|\theta)$ is sample model, the data model, or the likelihood function.
- $p(\theta)$ is the prior distribution of the parameter θ.
- $p(y)$ is the probability of data or evidence.
- $p(\theta|y)$ is the posterior distribution of the parameter given data.

Why Bayes?

- Probabilities can numerically represent a set of rational beliefs (i.e., fundamentally sound and based on rational rules).
- Explicit relationship between probability and information.
- Quantifies change in beliefs of a rational person when given new information (i.e., uses all available information–past & present).
- Very flexible
- Common sense interpretations.

In other words, if $p(\theta)$ approximates our beliefs, then $p(\theta|y)$ is optimal to what our posterior (after we have new information) beliefs about θ should be. Bayes can be used to explore how beliefs should be up-dated given data by someone with no information (e.g., who will win the 2020 presidental election?) or with some information (e.g., will is rain in Champaign in 2019?).

General Uses of a Bayesian Approach

- Parameter estimates with good statistical properties
- Parsimonious descriptions of observed data.
- Predictions for missing data.
- Predictions of future data.
- Computational frame-work for model estimation and validation.
- Provides a solution to complicated statistical problems that have no obvious (non-Bayesian) method of estimation and inference (e.g., complex statistical model, estimation of rare events).

Major Problems using Bayesian Approach

- Specifying prior knowledge; that is, choosing a prior.
- Sample from
$$\frac{p(y|\theta)p(\theta)}{p(y)}$$
- Computationally intensive

What do we Mean by “Probability”

Different authors of Bayesian texts use different terms for probability, which reflect different conceptualizations.

- Beliefs
- Credibility
- Plausibilities
- Subjective

There are multiple specific definitions:

- Frequentist: long run relative frequency of an event.
- Bayesian: a fundamental measure of uncertainty that follow rules probability theory.
  - What is the probability of thunder snow tomorrow?
  - What is the probability that Clinton nuclear power plant has a melt down?
  - What is the probability that a coin tossed lands on head?

Probabilities as We've Known Them

Probabilities are foundational concept!
Probabilitiesare numerical quantities that measures of uncertainty.

Justification for the statement "The probability that an even number comes up on a toss of a dice equals 1/2."

Symmetry or exchangeability argument:

$$p(\text{even}) = \frac{\text{number of evens rolled}}{\text{number of possible results}}$$

The justification is based on the physical process of rolling a dice where we assume each side of a 6 sided die are equal likely, three sides have even numbers, the other three have odd numbers.

$y_1 =$even or odd on first roll should be the same as y_2 on 2nd, etc.

Probabilities as We've Known Them

Alternative justification for the statement "The probability that an even number comes up on a toss of a dice equals 1/2."

Frequency argument:

$$p(\text{even}) = \text{Long run relative frequency}$$

Long (infinite) sequence of physically independent rolls of the dice.

Are these justifications subjective? These involve hypotheticals: physical independence, infinite sequence of rolls, equally likely, mathematical idealizations.

How would either of these justifications apply to

- If we only roll dice once?
- What's the probability that USA womens soccer team wins the next World Cup?

Probabilities as measure of Uncertainty

- Randomness creates uncertainty and we already do it in common speech...what are synonyms for "probability"?
- Coherence: probabilities principles of basic axioms of probability theory, which have a consequences things such as :
  - $0 \leq \text{p(X)} \leq 1$
  - if X is subset or equal to Y, then $p(X) \leq p(Y)$
  - $\sum p(X) = 1$ or $\int p(X) = 1$ item $p(X, Y) = p(X) + p(Y) - p(X \bigcap Y)$

Expected Value

- Expected value is a mean of some statistic or quantity based on a random event or outcome.
- For discrete random variables, say X with "probability mass function" $p(x)$, the mean of X is

$$E(X) = \mu_X = \sum_{i=1}^{I} x_i Pr(X = x_i),$$

  where I is the number of possible values for x, and $Pr(X = x_i)$ is the probability that $X = x_i$.
- For continuous random variables, say X with a probability density function $f(x)$, the mean of X is

$$E(X) = \mu_X = \int_x xf(x)d(x),$$

  where integration is over all possible values of x.

Basic Steps of Bayesian Analysis

(from Gelman et al.)

I assume that you have research questions, collected relavent data, and know the nature of the data.

- Set up full probability model (a joint probability distribution for all observed and unobserved variables that reflect knowledge and how data were collected):

  $$p(y,\theta) = p(y|\theta)p(\theta) = p(\theta|y)p(y)$$

  This can be the hard part
- Condition on data to obtain the posterior distribution:

  $$p(\theta|y) = p(y|\theta)p(\theta)/p(y)$$

  Tools: analytic, grid approximation, Markov chain Monte Carlo (Metropolis-Hastings, Gibbs sampling, Hamiltonian).
- Evaluate model fit.

A Closer Look at Bayes Rule

I am being vagues in terms of what y and θ are (e.g., continuous, discrete, number of parameters, what the data are).

$$\begin{aligned} p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\ &\propto p(y|\theta)p(\theta) \end{aligned}$$

where $p(y)$

- ensures probability sums to 1.
- is constant (for a given problem).
- "average" of numerator or the evidence.
- For discrete y: $p(y) = \sum_{\theta\in\Theta} p(y|\theta)p(\theta)$
- For continuous y: $p(y) = \int_\theta p(y|\theta)p(\theta)d(\theta)$

More on the Uses of a Bayesian Approach

- If $p(\theta)$ is wrong and doesn't represent our prior beliefs, the posterior is still useful. The posterior, $p(\theta|y)$, is optimal under $p(\theta)$ which means that $p(\theta|y)$ will generally serve as a good approximation of what our beliefs should be once we have data.
- Can use Bayesian approach to investigate how data would be updated using (prior) beliefs from different people. You can look at how opinions may changes for someone with *weak prior information* (vs someone with strong prior beliefs). Often diffuse or flat priors are used.
- Can handle complicated problems.

Example: Spell Checker

(from Gelman et al.)

Suppose the word "radom" is entered and we want to know the probability that this is the word intended, but there 2 other similar words that differ by one letter.

θ	frequency from Google database	Prior $p(\theta)$	Google's model $p(\text{'radom'}\|\theta)$	numerator $p(\theta)p(\text{'radom'}\|\theta)$
random	7.60×10^{-5}	0.9227556	0.001930	1.47×10^{-7}
radon	6.05×10^{-6}	0.0734562	0.000143	8.65×10^{-10}
radom	3.12×10^{-7}	0.0037881	0.975000	3.04×10^{-7}
total	8.2362×10^{-5}	1.00	1.00	4.51867×10^{-7}

Spell Checker (continued)

		Google's	numerator	
	Prior	model	of Bayes	Posterior
θ	p('radom')	p('radom'$\|\theta$)	$p(\theta)p$('radom'$\|\theta$)	$p(\theta\|$'radon')
random	0.9227556	0.001930	1.47×10^{-7}	0.325
radon	0.0734562	0.000143	8.65×10^{-10}	0.002
radom	0.0037881	0.975000	3.04×10^{-7}	0.673
total	1.00	1.00	4.51867×10^{-7}	1.000

Spell Checker (continued)

θ	Prior $p(\theta)$	Data model $p(\text{'radom'}\mid\theta)$	Posterior $p(\theta\mid\text{'radon'})$
random	0.9227556	0.001930	0.325
radon	0.0734562	0.000143	0.002
radom	0.0037881	0.975000	0.673
total	1.00	1.00	1.000

What is "radom"?

Some averaging of prior and data going on...most in this next lecture.

What could be some criticisms of this example or how might it be improved?

History: Rev Thomas Bayes

sources: Leonard (20140, Fienberg, S. (2006).

1700s

Keep in mind that statistics is a relatively newer field (only about 250 years old).

- 1763 Rev Thomas Bayes gave the first description of the theorem in “An essay toward solving a problem in the doctrine of chance”. This published posthumously by Richard Price.
- Bayes dealt with the problem of drawing inference; that is, concerned with “degree of probability”.
- Bayes introduces uniform prior distribution for binomial proportion.
- Price added an appendix that deals with the problem of prediction.
- Bayes did not give statement of what we call “Bayes Theorem”.
- 1749 David Hartley’s book describes the “inverse” result and attributes is to a friend. Speculation is that the friend was either Saunderson or Bayes.

1700s (continued)

- 1774 Pierre Simon LaPlace gave more elaborate version of Bayes theorem for the problem of inference for an unknown binomial probability in more modern language. He clearly augured for choosing a uniform prior because he reasoned that the posterior distribution of the probability should be proportional to the prior,

  $$f(\theta|x_1, x_2, \ldots x_n) \propto f(x_1, x_2, \ldots, x_n|\theta)$$

  I think this is why the term "inverse probability" was used.
- LaPlace introduced the idea of "indifference" as an argument to use uniform prior; that is, you have no information what the parameter should be.
- I.J. Bienayme generalized LaPlace's work.
- von Mise gave a rigorous proof of Bayes theorem.

1800s

1837–1843: at least 6 authors, working independently, made distinctions between probabilities of things (objective) and subjective meaning of probability (i.e., S.D. Poisson, D. Bolzano, R.L Ellis, J.F. Frees, J.S. Mills and A.A. Counot).

Debate on meaning of probability continued throughout the 1800s.

Some adopted the inverse probability (i.e, Bayesian) but also argued for a role of experience, including Pearson, Gosset and others.

1900s

- 1912–1922: Fisher advocated moving away from inverse methods toward inference based on likelihood.
- Fished moved away from "inverse probability" and argued for a frequentist approach.
  "...the theory of inverse probability is founded upon an error, and must wholly be rejected."
- Fundamental change in thinking.
- Beginnings of formal methodology for significance tests.
- J Neyman & Ego Pearson gave more mathematical detail and extended ("completed") Fisher's work which gave rising to the hypohthesis test and confidence intervals

1900s (continued)

- After WWWI, frequentist methods usurped inverse probability and Bayesian statistician were marginalized.
- R. von Mises justified the frequentist notion of probability; however, in 1941 he used a Bayesian argument to critique Neyman's method for confidence intervals. He argued that what really is wanted was posterior distribution.
- 1940 Wald showed that Bayesian approach yielded good frequentist properties and helped to rescue Bayes Theorem from obscurity.
- 1950s The term "frequentist" starts to be used. The term "Bayes" or "Bayes solution" was already in use. The term "classical" statistics refers to the frequentist.

1900s (continued)

- J.M Keynes (1920s) laid out axiomatic formulation and new approach to subjective probabilities via the concept of expect utility. Some quotes that reflect this thinking:
  - "In the long run we are all dead."
  - "It is better to be roughly right than precisely wrong."
  - "When the facts change, I change my mind."
- 1930s: Bruno de Finetti gave a different justification for subject probabilities and introduced the notion of "exchangeability" and implicit role of the prior distribution.
- Savage build on de Finetti's ideas and developed set of axioms for non-frequentist probabilities.

WWWII

- Alan Turing and his code breaking work was essentially Bayesian — sequential data analysis using weights of evidence. It is thought that he independently thought of these ideas.
- Decision-theory developments in the 1950s.

1980s and Beyond

Large revival started in the late 1980s and 1990s. This was due to new conceptual appoarches and lead to rapid increases in applications. The increase in computing power helped fuel this.

Non-Bayesian approaches will likely remain important because of the hight computational demand and expense of Bayeisan methods, even though there are continual developments in computing power.

Practice 1: Subjective Probability

Discuss the following statements:“The probability of event E is considered ‘subjective’ if two rational people A and B can assign unequal probabilities to E, $P_A(E)$ and $P_B(E)$. These probabilities can also be interpreted as ‘conditional’: $P_A(E) = P(E|I_A)$ and $P_B(E|I_A)$, where I_A and I_B represent the knowledge available to person A and B, respectively.” Apply this idea to the following examples.

- The probability that a “6” appears when a fair die is rolled, where A observes the outcome and B does not.
- The probability that USA wins the next mens World Cup, where A is ignorant of soccer and B is a knowledgable sports fan.
- The probability that UofI’s football team goes to a bowl game, where A is ignorant of Illini football and B is knowledgable of Illini football.

Practice 2: Conditional Probabilities and a little R

(from Gelman et al.)

Suppose that $\theta = 1$, then Y has a normal distribution with mean 1 and standard deviation σ, and if $\theta = 2$ then Y is normal with mean 2 and standard deviation σ. Also suppose $Pr(\theta = 1) = 0.5$ and $Pr(\theta = 2) = 0.5$.

- For $\sigma = 2$, write the formula for the marginal probability density for y and sketch/plot it. For the graphr, you will need to use the R commands:
  - *seq*
  - *dnorm*
  - *plot*
- What is $Pr(\theta = 1|y = 1)$ and what is $Pr(\theta = 1|y = 2)$. (hint: Definition of conditional probability, Bayes Theorem)
- Describe how the posterior density of θ in shape as
  - σ increases
  - σ decreases
  - Difference between μ's increase.
  - Different between μ's decrease.