Bayesian Analysis for the Social Sciences

Simon Jackman

Stanford University
http://jackman.stanford.edu/BASS

February 4, 2012

Introduction to Bayesian Inference

- Bayesian inference relies exclusively on Bayes Theorem:

$$\boxed{p(\theta|\text{data}) \propto p(\theta)\, p(\text{data}|\theta)}$$

- θ is a usually a parameter (but could also be a data point, a model, a hypothesis)
- p are **probability densities** (or probability mass functions in the case of discrete θ and/or discrete data)
- $p(\theta)$ a **prior** density; $p(\text{data}|\theta)$ the **likelihood** or **conditional density** of the data given θ
- $p(\theta|\text{data})$ is the **posterior** density for θ given the data.
- Gives rise to the ***Bayesian mantra***:

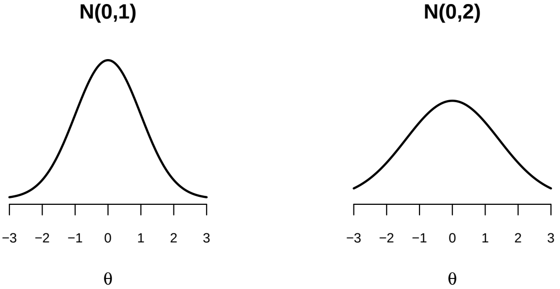
a posterior density is proportional to the prior times the likelihood

Probability Densities as Representations of Beliefs

Definition (Probability Density Function (informal))

Let θ be a unknown quantity, $\theta \in \Theta \subseteq \mathbb{R}$. A function $p(\theta)$ is a proper probability density function if

1. $p(\theta) \geq 0 \, \forall \; \theta$.
2. $\int_\Theta p(\theta) d\theta = 1$.

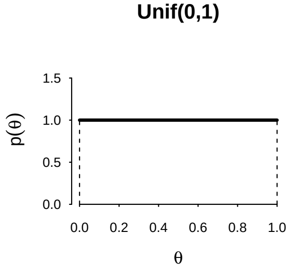


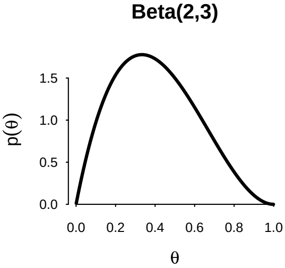
Probability Densities as Representations of Beliefs

Definition (Probability Density Function (informal))

Let θ be a unknown quantity, $\theta \in \Theta \subseteq \mathbb{R}$. A function $p(\theta)$ is a proper probability density function if

1. $p(\theta) \geq 0 \, \forall \; \theta.$
2. $\int_{\Theta} p(\theta) d\theta = 1.$



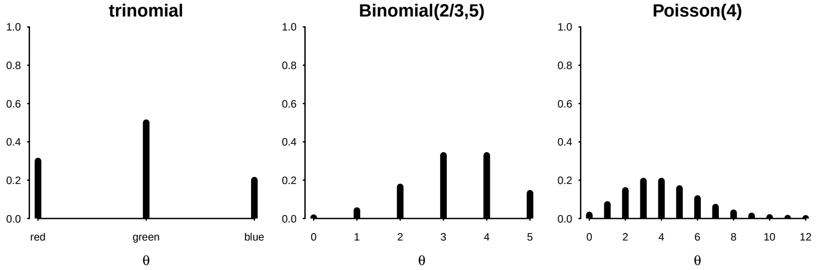


Probability Mass Function

Definition (Probability Mass Function)

If θ is a discrete random variable, taking values in a countable space $\Theta \subset \mathbb{R}$, then a function $p : \Theta \mapsto [0, 1]$ is a probability mass function if

1. $p(\theta) = 0 \,\forall\, \theta \in \mathbb{R} \setminus \Theta$
2. $\sum_{\theta \in \Theta} p(\theta) = 1$



Introduction to Bayesian Inference

$$\boxed{p(\theta|\text{data}) \propto p(\theta)\, p(\text{data}|\theta)}$$

- Bayesian inference involves **computing**, **summarizing** and **communicating** summaries of the **posterior density** $p(\theta|\text{data})$.
- How to do this is **what this class is about.**
- Depending on the problem, doing all this is easy or hard; we solve "hard" with computing power.
- We're working with densities (or sometimes, mass functions).
- Bayesian point estimates are a single number summary of a posterior density
- Uncertainty assessed/communicated in various ways: e.g., the standard deviation of the posterior, width of interval spanning 2.5th to 97.5th percentiles of the posterior, etc.
- Sometimes, can just draw a picture; details, examples coming.

Introduction to Bayesian Inference

$$\boxed{p(\theta|\text{data}) \propto p(\theta)\, p(\text{data}|\theta)}$$

- Bayes Theorem tells us how to *update* beliefs about θ in light of evidence (“data”)
- a general method for ***induction*** or for “learning from data”:

$$\text{prior} \longrightarrow \text{data} \longrightarrow \text{posterior}$$

- Bayes Theorem is itself uncontroversial: follows from widely accepted axioms of probability theory (e.g., Kolmogorov) and the definition of conditional probability

Why Be Bayesian?

- **conceptual simplicity:** “say what you mean” and “mean what you say” (subjective probability)
- a foundation for inference that does not rest on the thought experiment of repeated sampling
- **uniformity of application:** no special tweeks for this or that data analysis. Apply Bayes Rule.
- **modern computing** makes Bayesian inference easy and nearly universally applicable

Conceptual Simplicity

$$\boxed{p(\theta|\text{data}) \propto p(\theta)\, p(\text{data}|\theta)}$$

- the posterior density (or mass function) $p(\theta|\text{data})$ is a complete characterization of beliefs after looking at data
- as such it contains everything we need for making inferences
- Examples:
  - the posterior probability that a regression coefficient is positive, negative or lies in a particular interval;
  - the posterior probability that a subject belongs to a particular latent class;
  - the posterior probability that a hypothesis is true; or,
  - the posterior probabilities that a particular statistical model is true model among a family of statistical models.

Contrast Frequentist Inference

- Model for data: $y \sim f(\theta)$.
- Estimate θ: e.g., least squares, MLE, etc, to yield $\hat{\theta} \equiv \hat{\theta}(y)$
- null hypothesis e.g., $H_0 : \theta_{H_0} = 0$
- Inference via the *sampling distribution* of $\hat{\theta}$ conditional on H_0: e.g.,

> assuming H_0, over repeated applications of the sampling process, *how frequently* would we observe a result *at least as extreme* as the one we obtained?

- "At least as extreme"? Assessed via a test statistic, e.g.,

$$t(y) = (\theta_{H_0} - \hat{\theta}) / \sqrt{\text{var}(\hat{\theta} | \theta = \theta_{H_0})}$$

- "how frequently"? The p-value, relative frequency with which we see $|t| > t(y)$ in *repeated applications of the sampling process*. Often $t(y) \xrightarrow{d} N(0, 1)$.

Contrast Frequentist Inference

- null hypothesis e.g., $H_0 : \theta_{H_0} = 0$
- test-statistic:

$$t(y) = (\theta_{H_0} \text{ - } \hat{\theta}) / \sqrt{\text{var}(\hat{\theta} | \theta = \theta_{H_0})}$$

- Often $t(y) \xrightarrow{d} N(0, 1)$.
- p-value is a statement about the plausibility of the statistic $\hat{\theta}$ relative to what we might have observed in random sampling assuming $H_0 : \theta_{H_0} = 0$
- one more step need to reject/fail-to-reject H_0. Is p sufficiently small?
- frequentist p-value is a summary of the distribution of $\hat{\theta}$ under H_0

Contrast Frequentist Inference

- n.b., frequentist inference treats $\hat{\theta}$ as a random variable
- θ is a fixed but unknown feature of the population from which data is being (randomly) sampled
- Bayesian inference: $\hat{\theta}$ is fixed, a function of the data available for analysis
- Bayesian inference: θ is a random variable, subject to (subjective) uncertainty

	Bayesian	Frequentist
θ	random	fixed but unknown
$\hat{\theta}$	fixed	random
"random-ness"	subjective	sampling
distribution of interest	posterior	sampling distribution
	$p(\theta \vert y)$	$p(\hat{\theta}(y) \vert \theta = \theta_{H_0})$

Subjective Uncertainty

- how do we do statistical inference in situations where repeated sampling is infeasible?
- inference when we have the entire population and hence no uncertainty due to sampling: e.g., parts of comparative political economy.
- Bayesians rely on a notion of *subjective* uncertainty
- e.g., θ is a random variable because we don't know its value
- Bayes Theorem tells us how to manage that uncertainty, how to update beliefs about θ in light of data
- Contrast objectivist notion of probability: probability as a property of the object under study (e.g., coins, decks of cards, roulette wheels, people, groups, societies).

Subjective Uncertainty

Many Bayesians regard objectivist probability as **metaphysical nonsense**. de Finetti:

> *PROBABILITY DOES NOT EXIST*
>
> *The abandonment of superstitious beliefs about...Fairies and Witches was an essential step along the road to scientific thinking. Probability, too, if regarded as something endowed with some kind of objective existence, is not less a misleading misconception, an illusory attempt to exteriorize or materialize our true probabilistic beliefs. In investigating the reasonableness of our own modes of thought and behaviour under uncertainty, all we require, and all that we are reasonably entitled to, is consistency among these beliefs, and their reasonable relation to any kind of relevant objective data ("relevant" in as much as subjectively deemed to be so). This is Probability Theory.*

Subjective Uncertainty

- Bayesian probability statements are thus about states of mind over states of the world, and not about states of the world *per sé*.
- Borel: one can guess the outcome of a coin toss while the coin is still in the air and its movement is perfectly determined, or even after the coin has landed but before one reviews the result.
- i.e., subjective uncertainty obtains irrespective of "objective uncertainty (however conceived)"
- not just any subjective uncertainty: beliefs must conform to the rules of probability: e.g., $p(\theta)$ should be *proper*: i.e., $\int_\Theta p(\theta) d\theta = 1$, $p(\theta) \geq 0 \, \forall \, \theta \in \Theta$.

Bayes Theorem

- Conditional probability: Let A and B be events with $P(B) > 0$. Then the conditional probability of A given B is

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A, B)}{P(B)}.$$

- Multiplication rule:

$$P(A \cap B) = P(A, B) = P(A|B)P(B) = P(B|A)P(A)$$

- Law of Total Probability:

$$P(B) = P(A \cap B) + P(\sim A \cap B) = P(B|A)P(A) + P(B| \sim A)P(\sim A)$$

- Bayes Theorem: If A and B are events with $P(B) > 0$, then

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes Theorem, Example case, drug-testing

- Prior work suggests that about 3% of the subject pool (elite athletes) uses a particular prohibited drug.
- H_U: test subject uses the prohibited substance.
- $p(H_U) = .03$.
- E (evidence) is a positive test result.
- Test has a false negative rate of .05; i.e., $P(\sim E|H_U) = .05 \Rightarrow P(E|H_U) = .95$.
- Test has a false positive rate of .10: i.e., $P(E|H_{\sim U}) = .10$.
- Bayes Theorem:

$$\begin{aligned} P(H_U|E) &= \frac{P(H_U)P(E|H_U)}{\sum_{i \in \{U, \sim U\}} P(H_i)P(E|H_i)} \\ &= \frac{.03 \times .95}{(.03 \times .95) + (.97 \times .10)} \\ &= \frac{.0285}{.0285 + .097} \\ &= .23 \end{aligned}$$

Bayes Theorem, Continuous Parameter

- Bayes Theorem:

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

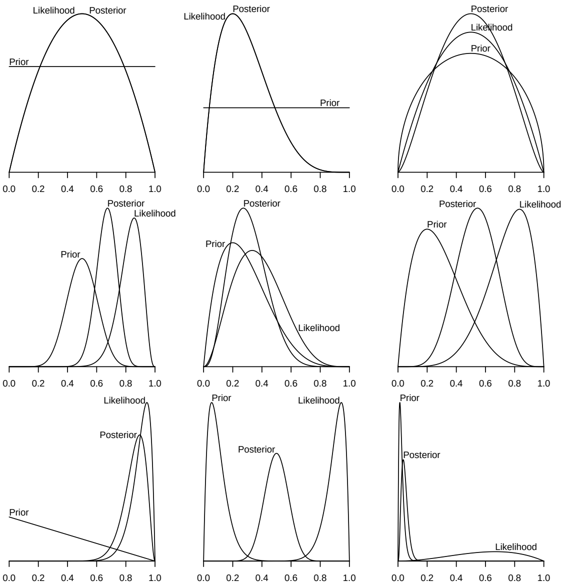
- Proof: by the definition of conditional probability

$$p(\theta, y) = p(\theta|y)p(y) = p(y|\theta)p(\theta), \tag{1}$$

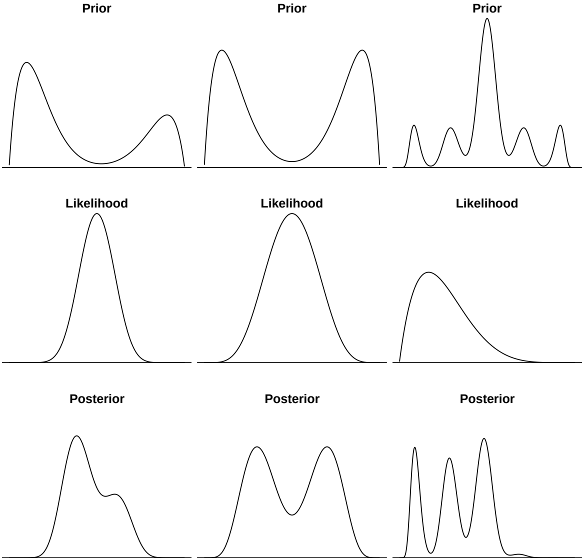
  where all these densities are assumed to exist and have the properties $p(z) > 0$ and $\int p(z)dz = 1$ (i.e., are *proper* probability densities.

  The result follows by re-arranging the quantities in equation equation 1 and noting that $p(y) = \int p(y, \theta)d\theta = \int p(y|\theta)p(\theta)d\theta$.

Prior and Posterior Densities, Continuous Parameter



Prior, Likelihood and Posteriors: less standard cases

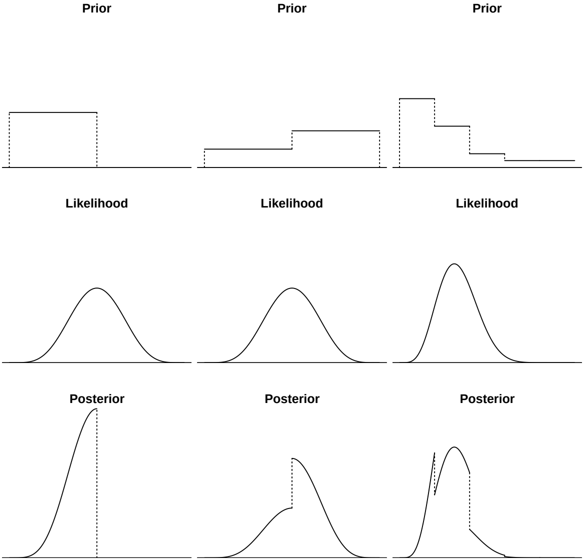


Cromwell's Rule: the dangers of dogmatism

$$\boxed{p(\theta|\text{data}) \propto p(\theta)\, p(\text{data}|\theta)}$$

- $p(\theta|\text{data}) = 0 \forall\, \theta \text{s.t.} p(\theta) = 0.$
- Cromwell's Rule: After the English deposed, tried and executed Charles I in 1649, the Scots invited Charles' son, Charles II, to become king. The English regarded this as a hostile act, and Oliver Cromwell led an army north. Prior to the outbreak of hostilities, Cromwell wrote to the synod of the Church of Scotland, "I beseech you, in the bowels of Christ, consider it possible that you are mistaken".
- a dogmatic prior that assigns zero probability to a hypothesis can never be revised
- likewise, a hypothesis with prior weight of 1.0 can never be refuted.

Cromwell's Rule



Bayesian Point Estimates

- ***Bayes estimates***: single number summary of a posterior density
- but which one?: e.g., mode, median, mean, some quantile(s)?
- different loss functions rationalize different point estimate
- Loss: Let Θ be a set of possible states of nature θ, and let $a \in \mathcal{A}$ be actions availble to the researcher. Then define $l(\theta, a)$ as the *loss* to the researcher from taking action a when the state of nature is θ.
- Posterior expected loss: Given a posterior distribution for θ, $p(\theta|\mathbf{y})$, the posterior expected loss of an action a is $v(p(\theta|\mathbf{y}), a) = \int_\Theta l(\theta, a) p(\theta|\mathbf{y}) d\theta$.

Posterior Mean as Bayes Estimator Under Quadratic Loss

- quadratic loss: If $\theta \in \Theta$ is a parameter of interest, and $\tilde{\theta}$ is an estimate of θ, then $l(\theta, \tilde{\theta}) = (\theta - \tilde{\theta})^2$ is the quadratic loss arising from the use of the estimate $\tilde{\theta}$ instead of θ.
- Posterior Mean as Bayes Estimate Under Quadratic Loss:

$$\boxed{E(\theta|\mathbf{y}) = \tilde{\theta} = \int_\Theta \theta\, p(\theta|\mathbf{y}) d\theta.}$$

- Proof: Quadratic loss implies that the posterior expected loss is

$$v(\theta, \tilde{\theta}) = \int_\Theta (\theta - \tilde{\theta})^2 p(\theta|\mathbf{y}) d\theta.$$

Expanding the quadratic yields $v(\theta, \tilde{\theta}) = \int_\Theta \theta^2 p(\theta|\mathbf{y}) d\theta + \tilde{\theta}^2 - 2\tilde{\theta} E(\theta|\mathbf{y})$. Differentiate with respect to $\tilde{\theta}$, noting that the first term does not involve $\tilde{\theta}$. Solve the 1st order condition for $\tilde{\theta}$ and the result follows.

Bayes Estimates

- Quadratic Loss: mean of the posterior density,

$$E(\theta|y) = \int_{\Theta} \theta \, p(\theta|\mathbf{y}) d\theta$$

- Symmetric Linear Loss: median of the posterior density, n.b., only well-defined for $\theta \in \Theta \subseteq \mathbb{R}$, in which case $\tilde{\theta}$ is defined such that

$$\int_{=} .5$$

- All-or-nothing Loss: mode of the posterior density

$$\tilde{\theta} = \operatorname*{argmax}_{\theta \in \Theta} p(\theta|y)$$

Credible Region; HPD region

Definition (Credible Region)

A region $C \subseteq \Omega$ such that $\int_C p(\theta) d\theta = 1 - \alpha, \ 0 \leq \alpha \leq 1$ is a $100(1 - \alpha)\%$ credible region for θ.
For single-parameter problems (i.e., $\Omega \subseteq \mathbb{R}$), if C is not a set of disjoint intervals, then C is a credible interval.
If $p(\theta)$ is a (prior/posterior) density, then C is a (prior/posterior) credible region.

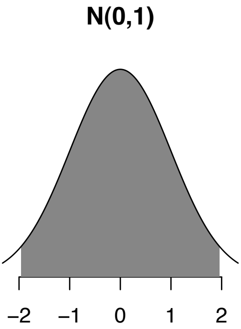
Definition (Highest Probability Density Region)

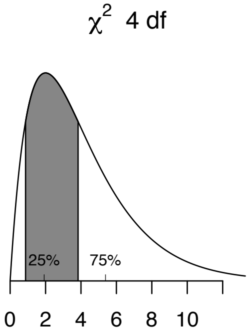
A region $C \subseteq \Omega$ is a $100(1 - \alpha)\%$ highest probability density region for θ under $p(\theta)$ if

1. $P(\theta \in C) = 1 - \alpha$
2. $P(\theta_1) \geq P(\theta_2), \forall \, \theta_1 \in C, \theta_2 \notin C$

HPD intervals

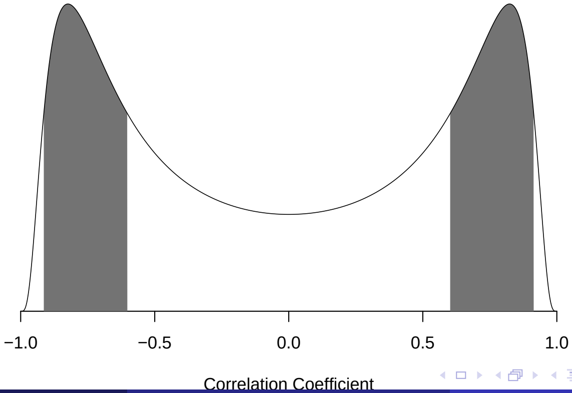
- A 100(1 - α)% HPD region for a symmetric, unimodal density is unique and symmetric around the mode; e.g., a normal density.
- Cf skewed distributions; a HPD differs from simply reading off the quantiles.





HPD intervals

- HPDs can be a series of disjoint intervals, e.g., a bimodal density
- these are uncommon; but in such a circumstance, presenting a picture of the density might be the reasonable thing to do.
- See Example 1.7, p28: $\mathbf{y}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, subject to extreme missingness. The posterior density of $\rho(\boldsymbol{\Sigma}) = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$:

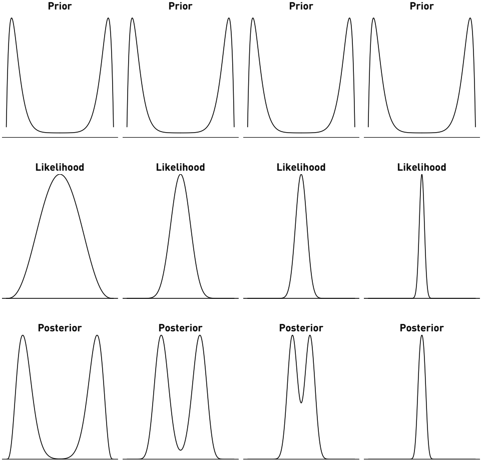


Bayesian Consistency

- for anything other than a dogmatic/degenerate prior (see the earlier discussion of Cromwell's Rule), more and more data will overwhelm the prior.
- Bayesian asymptotics: with an arbitrarily large amount of sample information relative to prior information, the posterior density tends to the likelihood (normalized to be a density over θ).
- central limit arguments: since likelihoods are usually approximately normal in large samples, then so too are posterior densities.

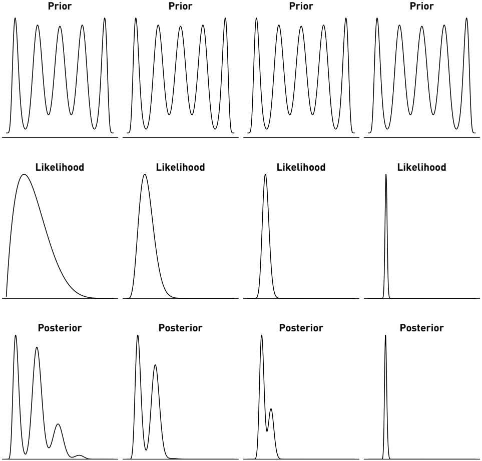
Bayesian Consistency

The prior remains fixed across the sequence, as sample size increases and θ^* is held constant. In this example, $n = 6, 30, 90, 450$ across the four columns.



Bayesian Consistency

The prior remains fixed across the sequence, as sample size increases and θ^* is held constant. In this example, $n = 6, 30, 150, 1500$ across the four columns.



Other topics from Chapter One

- §1.8. Bayesian hypothesis testing.
- §1.9. Exchangeability. de Finetti's Representation Theorem.