Introduction to Bayesian Inference

What is probability?

1. Axiomatic probability.
2. Classical probability: principle of indifference.
3. Frequentist probability: $Pr(A) = lim_{n \to \inf} \frac{m}{n}$
4. Subjective probability.

Subjective Probability

- Probability is a measure of the uncertainty that a given individual associates with a particular statement.
- Any statement can be assigned a probability.
- Can assign probabilities to any uncertain outcome.
- Coherent (non-exploitable) subjective beliefs follow the axioms of probability.

Subjective Probability

- Bayesian probability statements are about states of mind over states of the world, and not about states of the world *per se*.
- 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.
- Not just any subjective uncertainty: beliefs must conform to the rules of probability.

The Mantra

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

- A posterior density is proportional to the prior times the likelihood
- It is a general method for induction or to "learn from data."
- *prior* $\rightarrow$ *data* $\rightarrow$ *posterior*

Contrast Frequentist Inference

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

Bayes Theorem

Bayes Theorem for the probability of two events A and B with Pr(B) > 0 states that

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

where:

- $Pr(A|B)$ posterior probability of A given B.
- $Pr(B|A)$ probability of B given A (likelihood).
- $Pr(A)$ prior probability of A (unconditional).
- $Pr(B)$ probability of B (unconditional; normalizes the posterior probability to 1).

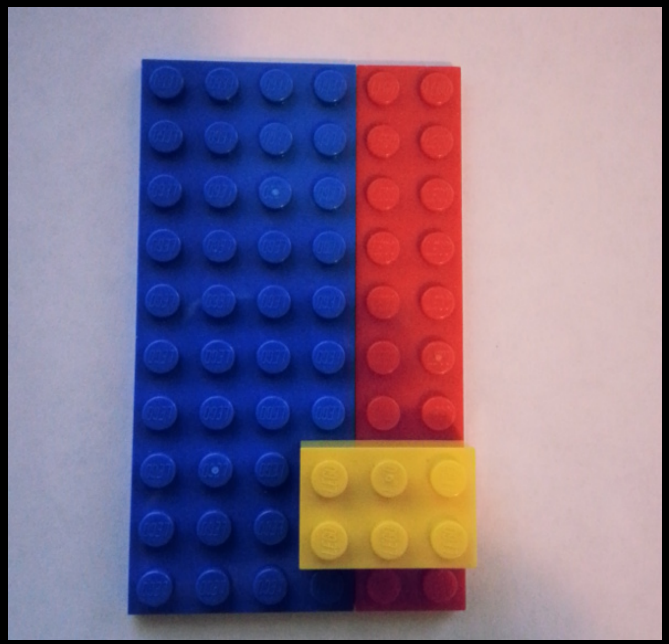
Example case

- Suppose a screening test can yield either a positive (P) or negative (N) result.
- Either result can obtain whether the patient is sick (S) or healthy (H).
- $Pr(P|S) = 0.99$
- $Pr(P|H) = 0.03$
- $Pr(S) = 0.05$

Example case

- We have to assess whether an athlete used a prohibited substance (U).
- 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.
- 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



Estimate the probability that we are on the yellow brick if the red brick is underneath

Bayes Theorem with random variables

- Let θ and y be random variables.
- If the distribution of θ is discrete:

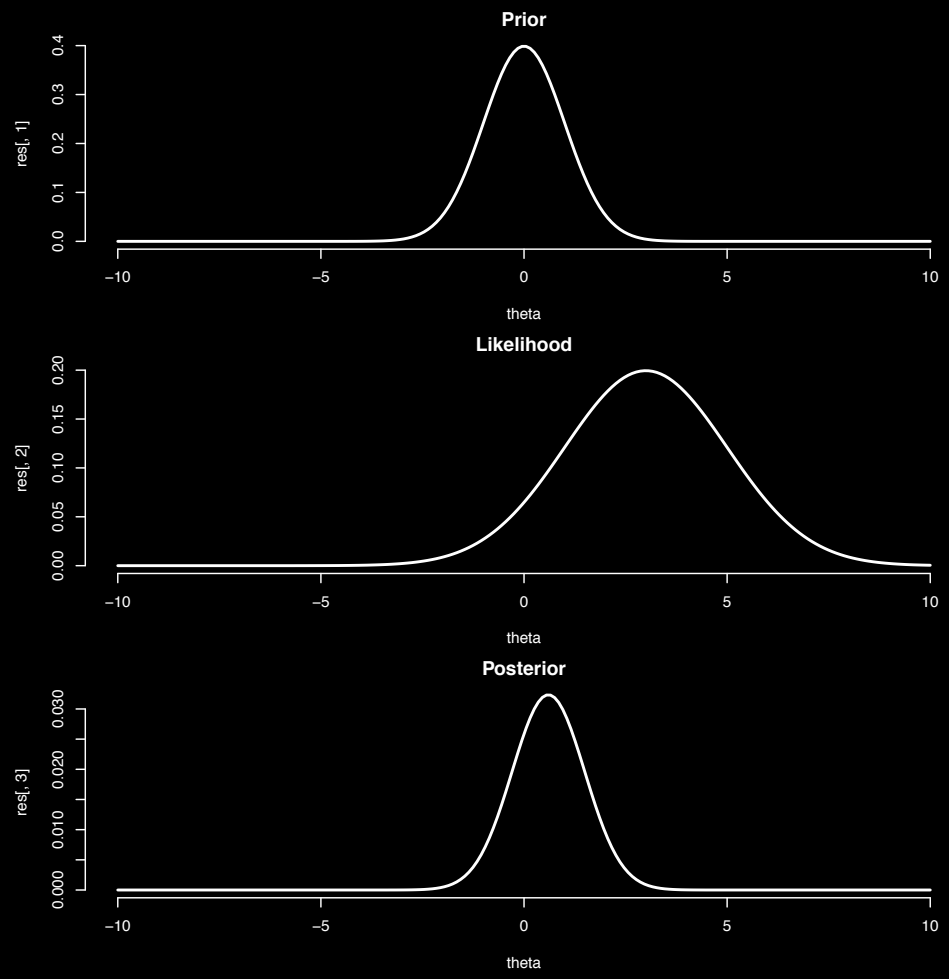
  $Pr(\theta|y) = \frac{Pr(y|\theta)Pr(\theta)}{\sum_{\theta \in S} p(y|\theta)p(\theta)}$, where S is the support of θ
- If the distribution of θ is continuous:

  $$Pr(\theta|y) = \frac{Pr(y|\theta)Pr(\theta)}{\int_{\theta \in S} p(y|\theta)p(\theta)d\theta}$$
- $Pr(B|A)$ probability of B given A (likelihood).

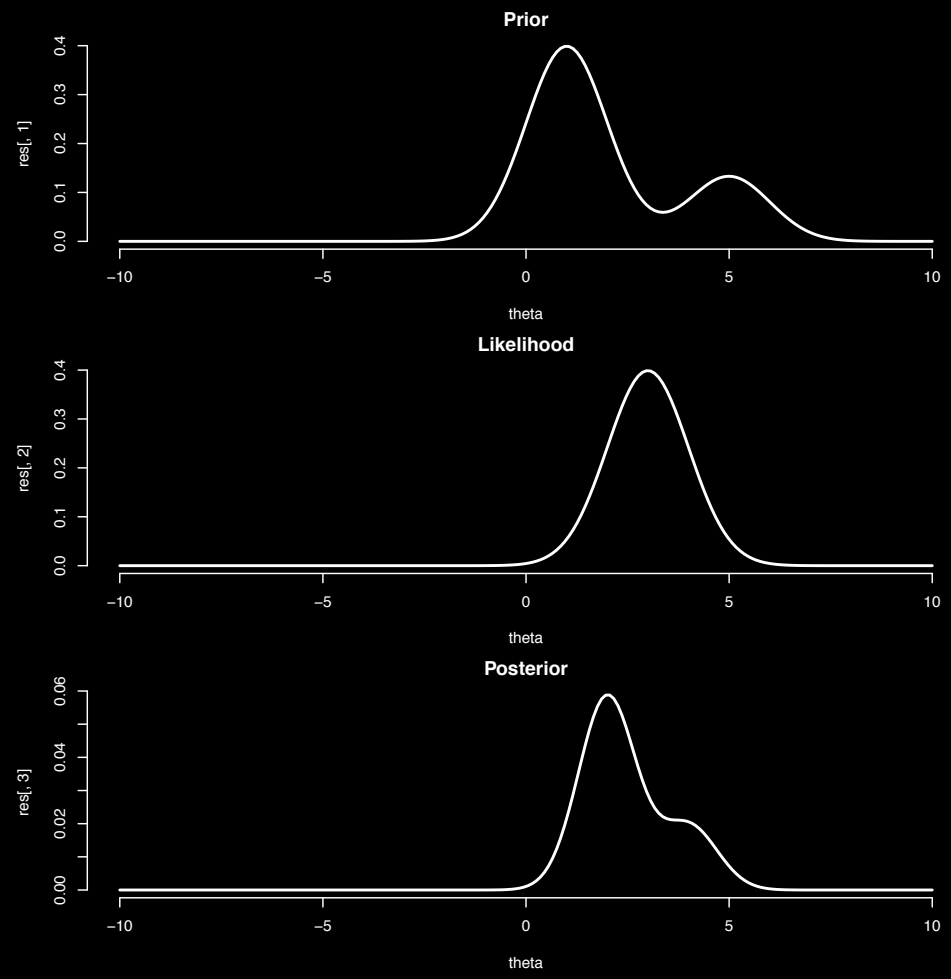
Bayes consistency

- 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.
- Thus Bayesian estimates converge to frequentist / *likelihoodist* estimates as $N \rightarrow$ inf.

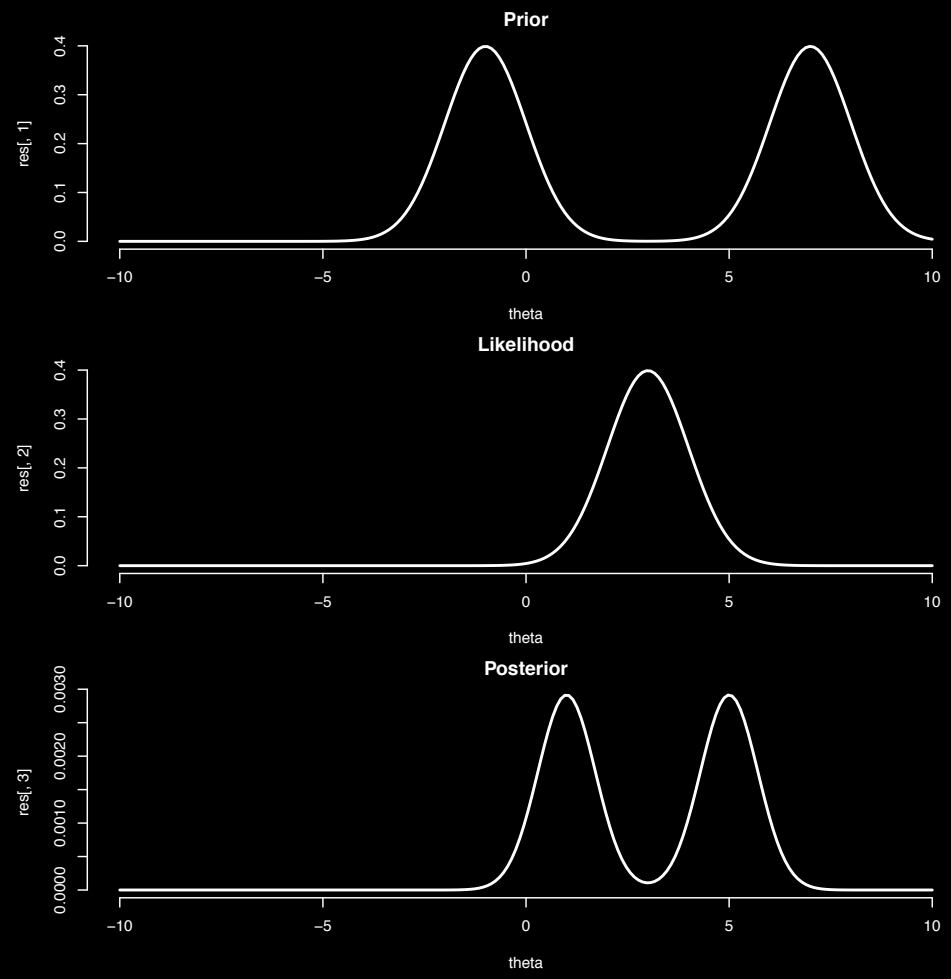
Prior, Likelihood, and Posteriors



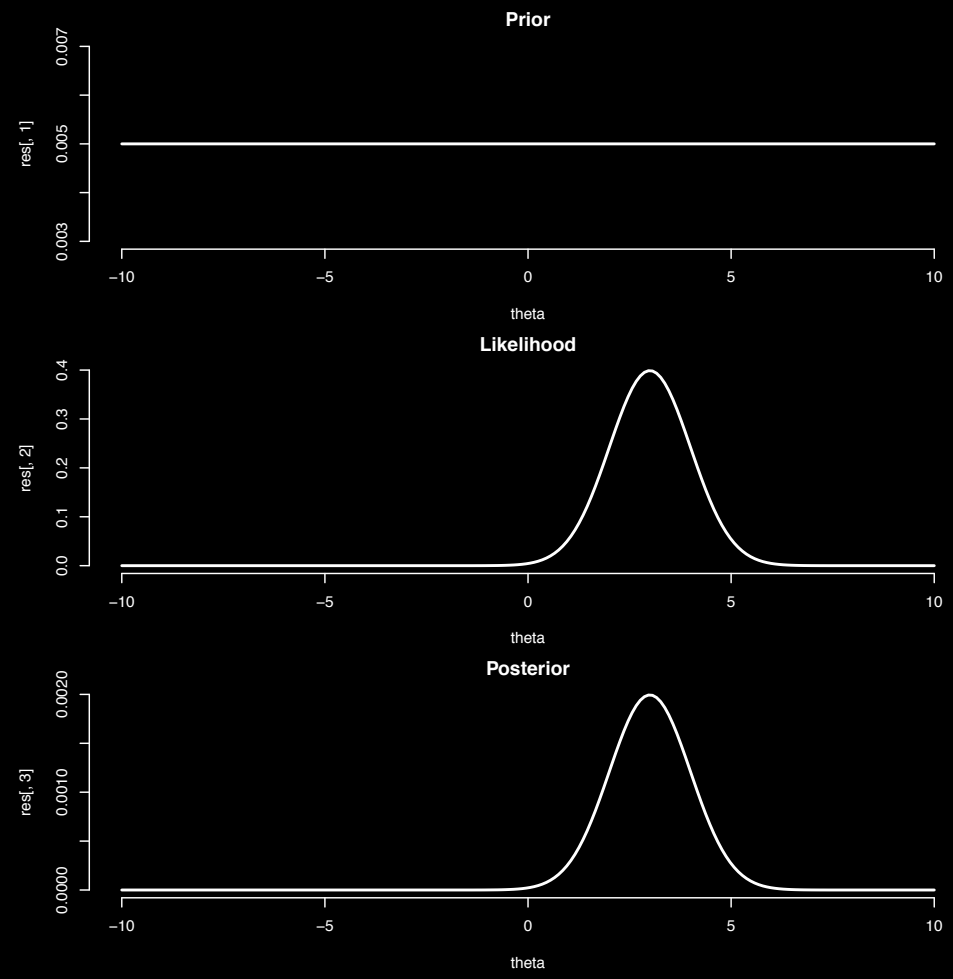
Prior, Likelihood, and Posteriors



Prior, Likelihood, and Posteriors



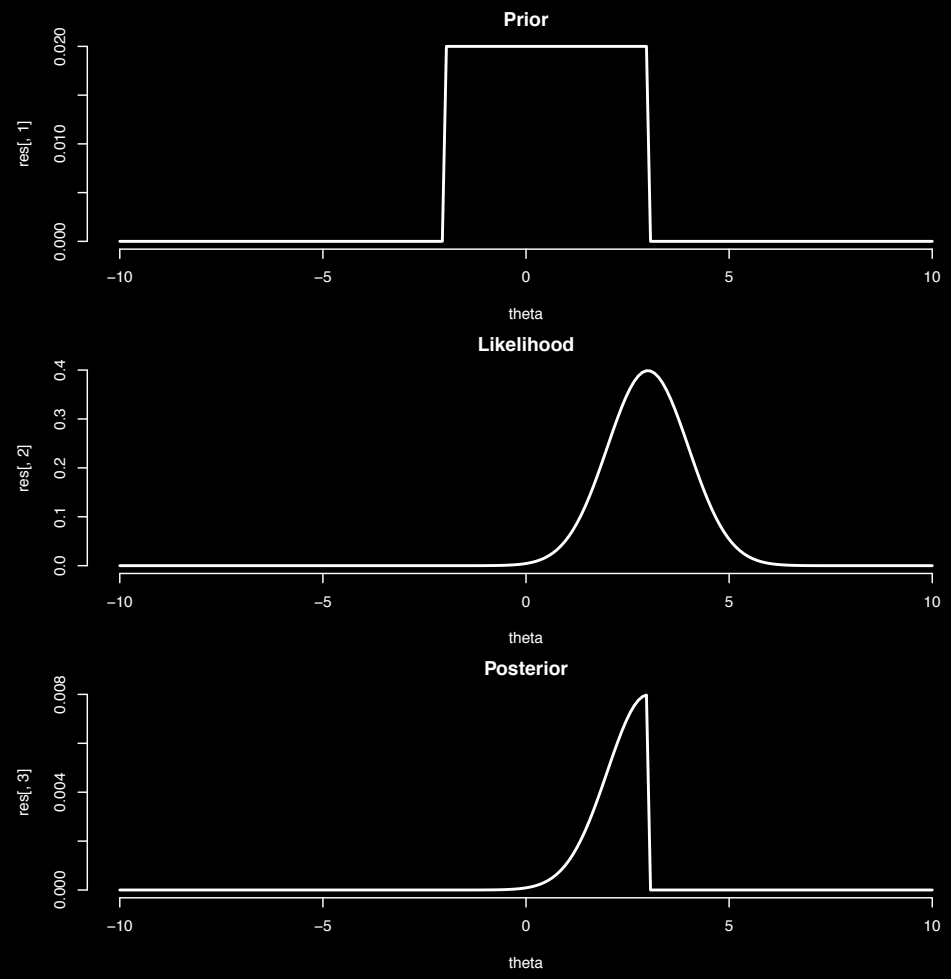
Prior, Likelihood, and Posteriors



Prior, Likelihood, and Posteriors

- Cromwell's Rule (or the dangers of dogmatism).
  - A dogmatic prior that assigns either zero or one probability to a hypothesis can never be revised
  - 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."

Prior, Likelihood, and Posteriors



Monte Carlo methods

- We will usually find problems that are hard to solve analytically.
- It is often possible to create a set of simulated values from a target distribution that share the same distributional properties even if we can't describe analytically, or even sample directly from, that distribution.
- Applied Bayesian statistics describes posterior beliefs using empirical summaries of the posterior distribution sampled via monte carlo methods.

Monte Carlo methods

- Example: suppose we are interested in the expected value $E(\theta)$.
- Analytically, we would compute $E(\theta) = \int_a^b \theta p(\theta) d\theta$. In some cases, the math is hard.
- Monte Carlo estimate:
  - Sample θ from $p(\theta)$, T times, $\theta^{(1)}, \theta^{(2)}, ..., \theta^{(T)}$, with T large.
  - Calculate

$$\widehat{E(\theta)} = \frac{1}{T \sum_{t=1} T\theta^{(t)}}$$

  - The larger the T, the closer $\widehat{E(\theta)}$ to $E(\theta)$.
  - Generalizes to vectors θ and to functions of θ ,$h(\theta)$.

Markov chain Monte Carlo (MCMC)

- In general, we can't simply sample from the posterior density $p(\theta|y)$ because
  - *theta* is a big object with many elements.
  - $p(\theta|y)$ is a complicated function, difficult to sample from.
- Methods to sample from p usually require us to give up the independence of the series of sampled values.
- A Markov chain is a stochastic process: the common analogy is to a particle moving at random through space.
- The particle's location at time t is a random movement away from where it was at $t-1$. Even though the movement is random, the chain is auto-correlated, because t is more likely to be close to $t-1$ than far.

Markov chain

- Consider a sequence of random variables (a stochastic process), $\theta^{(t)}$ for $t = 1, 2, 3, ...T$.
- We can write:

$$Pr(\theta^{(t+1)}|\theta^{(t)}, ..., \theta^{(1)})$$

- That is, we can characterize the probabilities of the stochastic outcome $\theta^{(t+1)}$ in terms of the prior history of the chain.
- If the process is a Markov chain, the next value depends only on the most recent value:

$$Pr(\theta^{(t+1)}|\theta^{(t)}, ..., \theta^{(1)}) = Pr(\theta^{(t+1)}|\theta^{(t)})$$

Applications

- House effects in electoral polls

$$y_{i,j} \sim Bin(\pi_{i,j}, n_i)$$
$$\pi_{i,t} = \alpha_{t[j]} + \delta_{k[j]}$$