Introduction to Bayesian Statistics

William Ruth

June 2022

What is a Parameter?

What is a Parameter?

- Target of inference

What is a Parameter?

- Target of inference
- To a Frequentist:

What is a Parameter?

- Target of inference
- To a Frequentist:
    - A fixed, unknown number that exists in the world

What is a Parameter?

- Target of inference
- To a Frequentist:
  - A fixed, unknown number that exists in the world
  - Study by repeated sampling

What is a Parameter?

- Target of inference
- To a Frequentist:
  - A fixed, unknown number that exists in the world
  - Study by repeated sampling
    - Usually hypothetical repeated sampling

What is a Parameter?

- Target of inference
- To a Frequentist:
  - A fixed, unknown number that exists in the world
  - Study by repeated sampling
    - Usually hypothetical repeated sampling
- To a Bayesian:

What is a Parameter?

- Target of inference
- To a Frequentist:
  - A fixed, unknown number that exists in the world
  - Study by repeated sampling
    - Usually hypothetical repeated sampling
- To a Bayesian:
  - An unknown number

What is a Parameter?

- Target of inference
- To a Frequentist:
  - A fixed, unknown number that exists in the world
  - Study by repeated sampling
    - Usually hypothetical repeated sampling
- To a Bayesian:
  - An unknown number
  - Quantify my beliefs about it

What is a Parameter?

- Target of inference
- To a Frequentist:
  - A fixed, unknown number that exists in the world
  - Study by repeated sampling
    - Usually hypothetical repeated sampling
- To a Bayesian:
  - An unknown number
  - Quantify my beliefs about it
  - Systematically update my beliefs using data

Quantifying and Updating Beliefs

Quantifying and Updating Beliefs

- Represent beliefs with probability distributions

Quantifying and Updating Beliefs

- Represent beliefs with probability distributions
  - E.g. I think it's twice as likely to rain tomorrow than not

Quantifying and Updating Beliefs

- Represent beliefs with probability distributions
  - E.g. I think it's twice as likely to rain tomorrow than not
  - E.g. I think that average rainfall in June in Vancouver is around 50mm and differing by more than 20mm is unlikely

Quantifying and Updating Beliefs

- Represent beliefs with probability distributions
  - E.g. I think it's twice as likely to rain tomorrow than not
  - E.g. I think that average rainfall in June in Vancouver is around 50mm and differing by more than 20mm is unlikely
    - Normal with mean of 50 and SD of 10

Quantifying and Updating Beliefs

- Represent beliefs with probability distributions
  - E.g. I think it's twice as likely to rain tomorrow than not
  - E.g. I think that average rainfall in June in Vancouver is around 50mm and differing by more than 20mm is unlikely
    - Normal with mean of 50 and SD of 10
- Much harder to do as a Frequentist

Quantifying and Updating Beliefs

- Represent beliefs with probability distributions
  - E.g. I think it's twice as likely to rain tomorrow than not
  - E.g. I think that average rainfall in June in Vancouver is around 50mm and differing by more than 20mm is unlikely
    - Normal with mean of 50 and SD of 10
- Much harder to do as a Frequentist
  - Need infinitely many days exactly like tomorrow

Quantifying and Updating Beliefs

- Represent beliefs with probability distributions
  - E.g. I think it's twice as likely to rain tomorrow than not
  - E.g. I think that average rainfall in June in Vancouver is around 50mm and differing by more than 20mm is unlikely
    - Normal with mean of 50 and SD of 10
- Much harder to do as a Frequentist
  - Need infinitely many days exactly like tomorrow
  - Or infinitely many Junes

Quantifying and Updating Beliefs

- Represent beliefs with probability distributions
  - E.g. I think it's twice as likely to rain tomorrow than not
  - E.g. I think that average rainfall in June in Vancouver is around 50mm and differing by more than 20mm is unlikely
    - Normal with mean of 50 and SD of 10
- Much harder to do as a Frequentist
  - Need infinitely many days exactly like tomorrow
  - Or infinitely many Junes
- Update probability distribution using data and Bayes Theorem

Conditional Probability

Conditional Probability

- Consider rolling a dice

Conditional Probability

- Consider rolling a dice
- What is probability of rolling a 2 given that you know the roll is even?

Conditional Probability

- Consider rolling a dice
- What is probability of rolling a 2 given that you know the roll is even?
  - Easy, $1/3$

Conditional Probability

- Consider rolling a dice
- What is probability of rolling a 2 given that you know the roll is even?
  - Easy, $1/3$
- Can we say something systematic here?

Conditional Probability

- Consider two possible outcomes, A and B

Conditional Probability

- Consider two possible outcomes, A and B
  - E.g. A is "roll a 2" and B is "roll even"

Conditional Probability

- Consider two possible outcomes, A and B
  - E.g. A is "roll a 2" and B is "roll even"
- Want probability of A given that we know B occurred

Conditional Probability

- Consider two possible outcomes, A and B
  - E.g. A is "roll a 2" and B is "roll even"
- Want probability of A given that we know B occurred
  - Write $P(A|B)$

Conditional Probability

- Consider two possible outcomes, A and B
  - E.g. A is "roll a 2" and B is "roll even"
- Want probability of A given that we know B occurred
  - Write $P(A|B)$
- Define

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

Conditional Probability

- Back to our example

Conditional Probability

- Back to our example

$$P(2|\text{even}) = \frac{P(2 \text{ and even})}{P(\text{even})}$$

Conditional Probability

- Back to our example

$$P(2|\text{even}) = \frac{P(2 \text{ and even})}{P(\text{even})}$$
$$= \frac{P(2)}{P(\text{even})}$$

Conditional Probability

- Back to our example

$$
\begin{aligned}
P(2|\text{even}) &= \frac{P(2 \text{ and even})}{P(\text{even})} \\
&= \frac{P(2)}{P(\text{even})} \\
&= \frac{1/6}{3/6}
\end{aligned}
$$

Conditional Probability

- Back to our example

$$
\begin{aligned}
P(2|\text{even}) &= \frac{P(2 \text{ and even})}{P(\text{even})} \\
&= \frac{P(2)}{P(\text{even})} \\
&= \frac{1/6}{3/6} \\
&= \frac{1}{3}
\end{aligned}
$$

Bayesian Terminology

Bayesian Terminology

- We start with a distribution for the unknown parameter

Bayesian Terminology

- We start with a distribution for the unknown parameter
  - Called the **prior distribution**

Bayesian Terminology

- We start with a distribution for the unknown parameter
  - Called the **prior distribution**
  - Denoted by $\pi(\theta)$

Bayesian Terminology

- We start with a distribution for the unknown parameter
  - Called the **prior distribution**
  - Denoted by $\pi(\theta)$
- We have some data, X

Bayesian Terminology

- We start with a distribution for the unknown parameter
  - Called the **prior distribution**
  - Denoted by $\pi(\theta)$
- We have some data, X
- Its distribution depends on the parameter, θ

Bayesian Terminology

- We start with a distribution for the unknown parameter
  - Called the **prior distribution**
  - Denoted by $\pi(\theta)$
- We have some data, X
- Its distribution depends on the parameter, θ
  - Distribution of the data, given the unknown parameter, is called the **likelihood**

Bayesian Terminology

- We start with a distribution for the unknown parameter
  - Called the **prior distribution**
  - Denoted by $\pi(\theta)$
- We have some data, X
- Its distribution depends on the parameter, θ
  - Distribution of the data, given the unknown parameter, is called the **likelihood**
  - Denoted by $L(X|\theta)$

Bayesian Terminology

- We would like to update the distribution of θ using observed data

Bayesian Terminology

- We would like to update the distribution of θ using observed data
  - I.e. Get the distribution of θ given X

Bayesian Terminology

- We would like to update the distribution of θ using observed data
  - I.e. Get the distribution of θ given X
- This is called the **posterior distribution**

Bayesian Terminology

- We would like to update the distribution of θ using observed data
  - I.e. Get the distribution of θ given X
- This is called the **posterior distribution**
  - Denoted by $\pi(\theta|X)$

Bayesian Terminology

- It can be shown that the posterior is proportional to the likelihood times the prior

Bayesian Terminology

- It can be shown that the posterior is proportional to the likelihood times the prior
  - I.e. $\pi(\theta|X) \propto L(X|\theta) \cdot \pi(\theta)$

Bayesian Terminology

- It can be shown that the posterior is proportional to the likelihood times the prior
  - I.e. $\pi(\theta|X) \propto L(X|\theta) \cdot \pi(\theta)$
- The proportionality constant depends on X but not on θ

Bayesian Terminology

- It can be shown that the posterior is proportional to the likelihood times the prior
  - I.e. $\pi(\theta|X) \propto L(X|\theta) \cdot \pi(\theta)$
- The proportionality constant depends on X but not on θ
- In principle, we can get the proportionality constant by integrating the posterior over the range of θ

Bayesian Terminology

- It can be shown that the posterior is proportional to the likelihood times the prior
  - I.e. $\pi(\theta|X) \propto L(X|\theta) \cdot \pi(\theta)$
- The proportionality constant depends on X but not on θ
- In principle, we can get the proportionality constant by integrating the posterior over the range of θ
  - Total probability must equal 1

Bayesian Terminology

- It can be shown that the posterior is proportional to the likelihood times the prior
  - I.e. $\pi(\theta|X) \propto L(X|\theta) \cdot \pi(\theta)$
- The proportionality constant depends on X but not on θ
- In principle, we can get the proportionality constant by integrating the posterior over the range of θ
  - Total probability must equal 1
- Usually, we can just ignore the constant

Bayesian Inference

Bayesian Inference

- In a sense, we're done

Bayesian Inference

- In a sense, we're done
- Anything you want to say about θ can be described in terms of the posterior

Bayesian Inference

- In a sense, we're done
- Anything you want to say about θ can be described in terms of the posterior

- Let's illustrate with an example

Example: Coin Tossing

Example: Coin Tossing

- Consider an experiment done at Berkley in 2009 in which a coin was tossed 40,000 times

Example: Coin Tossing

- Consider an experiment done at Berkley in 2009 in which a coin was tossed 40,000 times
  - So that we can actually do the calculations, we will just look at the first 100 flips

Example: Coin Tossing

- Consider an experiment done at Berkley in 2009 in which a coin was tossed 40,000 times
  - So that we can actually do the calculations, we will just look at the first 100 flips
  - Of the first 100 flips, 41 came up heads

Example: Coin Tossing

- Consider an experiment done at Berkley in 2009 in which a coin was tossed 40,000 times
  - So that we can actually do the calculations, we will just look at the first 100 flips
  - Of the first 100 flips, 41 came up heads
- Our parameter, θ, is the probability of getting heads

Example: Coin Tossing

- For the sake of illustration, let's use a uniform prior

Example: Coin Tossing

- For the sake of illustration, let's use a uniform prior
  - $\pi(\theta) = 1$ for $\theta \in [0, 1]$

Example: Coin Tossing

- For the sake of illustration, let's use a uniform prior
  - $\pi(\theta) = 1$ for $\theta \in [0, 1]$
- Given θ, the number of heads follows a binomial distribution

Example: Coin Tossing

- For the sake of illustration, let's use a uniform prior
  - $\pi(\theta) = 1$ for $\theta \in [0, 1]$
- Given θ, the number of heads follows a binomial distribution
  - $L(X|\theta) = \binom{100}{X} \cdot \theta^X \cdot (1-\theta)^{100-X}$

Example: Coin Tossing

- For the sake of illustration, let's use a uniform prior
  - $\pi(\theta) = 1$ for $\theta \in [0, 1]$

- Given θ, the number of heads follows a binomial distribution
  - $L(X|\theta) = \binom{100}{X} \cdot \theta^X \cdot (1-\theta)^{100-X}$
  - $L(X|\theta) \propto \theta^X \cdot (1-\theta)^{100-X}$

Example: Coin Tossing

- For the sake of illustration, let's use a uniform prior
  - $\pi(\theta) = 1$ for $\theta \in [0, 1]$

- Given θ, the number of heads follows a binomial distribution
  - $L(X|\theta) = \binom{100}{X} \cdot \theta^X \cdot (1-\theta)^{100-X}$
  - $L(X|\theta) \propto \theta^X \cdot (1-\theta)^{100-X}$

- Posterior is proportional to likelihood times prior

Example: Coin Tossing

- For the sake of illustration, let's use a uniform prior
  - $\pi(\theta) = 1$ for $\theta \in [0, 1]$

- Given θ, the number of heads follows a binomial distribution
  - $L(X|\theta) = \binom{100}{X} \cdot \theta^X \cdot (1-\theta)^{100-X}$
  - $L(X|\theta) \propto \theta^X \cdot (1-\theta)^{100-X}$

- Posterior is proportional to likelihood times prior
  - $\pi(\theta|X) \propto \theta^X \cdot (1-\theta)^{100-X} \cdot 1$

Example: Coin Tossing

- For the sake of illustration, let's use a uniform prior
  - $\pi(\theta) = 1$ for $\theta \in [0, 1]$

- Given θ, the number of heads follows a binomial distribution
  - $L(X|\theta) = \binom{100}{X} \cdot \theta^X \cdot (1-\theta)^{100-X}$
  - $L(X|\theta) \propto \theta^X \cdot (1-\theta)^{100-X}$

- Posterior is proportional to likelihood times prior
  - $\pi(\theta|X) \propto \theta^X \cdot (1-\theta)^{100-X} \cdot 1$
  - $\pi(\theta|X) = \theta^X \cdot (1-\theta)^{100-X}$

Example: Coin Tossing

- Up to proportionality constants, this posterior matches a beta distribution with parameters X and $100 - X$

Example: Coin Tossing

- Up to proportionality constants, this posterior matches a beta distribution with parameters X and $100 - X$
- Given data, θ follows a beta distribution with these parameter values

Example: Coin Tossing

- Up to proportionality constants, this posterior matches a beta distribution with parameters X and $100 - X$
- Given data, θ follows a beta distribution with these parameter values
  - We write $\theta | X \sim \text{Beta}(X, 100 - X)$

Example: Coin Tossing

- Up to proportionality constants, this posterior matches a beta distribution with parameters X and $100 - X$
- Given data, θ follows a beta distribution with these parameter values
  - We write $\theta|X \sim \text{Beta}(X, 100 - X)$
- Let's plug in our numbers

Example: Coin Tossing

- Up to proportionality constants, this posterior matches a beta distribution with parameters X and $100 - X$
- Given data, θ follows a beta distribution with these parameter values
  - We write $\theta | X \sim \text{Beta}(X, 100 - X)$
- Let's plug in our numbers
- X is 41, so $\theta | X \sim \text{Beta}(41, 59)$

Example: Coin Tossing

- The beta distribution is very well studied

Example: Coin Tossing

- The beta distribution is very well studied
- For a Beta(α, β) distribution,

Example: Coin Tossing

- The beta distribution is very well studied
- For a Beta(α, β) distribution,
  - The mean is $\frac{\alpha}{\beta+\alpha}$

Example: Coin Tossing

- The beta distribution is very well studied
- For a Beta(α, β) distribution,
  - The mean is $\frac{\alpha}{\beta+\alpha}$
  - The most likely value (mode) is $\frac{\alpha-1}{\alpha+\beta-2}$

Example: Coin Tossing

- The beta distribution is very well studied
- For a Beta(α, β) distribution,
  - The mean is $\frac{\alpha}{\beta+\alpha}$
  - The most likely value (mode) is $\frac{\alpha-1}{\alpha+\beta-2}$
- For our problem, mean is 0.41 and mode is 0.408

Example: Coin Tossing

- The beta distribution is very well studied
- For a Beta(α, β) distribution,
  - The mean is $\frac{\alpha}{\beta+\alpha}$
  - The most likely value (mode) is $\frac{\alpha-1}{\alpha+\beta-2}$
- For our problem, mean is 0.41 and mode is 0.408
- Unfortunately, we don't always get nice posteriors

Bayesian Computation

Bayesian Computation

- What if we started with a less trivial distribution for θ?

Bayesian Computation

- What if we started with a less trivial distribution for θ?
- Real world priors and likelihoods can get very complicated

Bayesian Computation

- What if we started with a less trivial distribution for θ?
- Real world priors and likelihoods can get very complicated

- The mean and mode on the previous slide are obtained analytically

Bayesian Computation

- What if we started with a less trivial distribution for θ?
- Real world priors and likelihoods can get very complicated

- The mean and mode on the previous slide are obtained analytically
- In general, we can't do the necessary integration or optimization

Bayesian Computation

- What if we started with a less trivial distribution for θ?
- Real world priors and likelihoods can get very complicated

- The mean and mode on the previous slide are obtained analytically
- In general, we can't do the necessary integration or optimization
- Instead, the posterior mean and posterior mode must be obtained numerically

Bayesian Computation

- What if we started with a less trivial distribution for θ?
- Real world priors and likelihoods can get very complicated

- The mean and mode on the previous slide are obtained analytically
- In general, we can't do the necessary integration or optimization
- Instead, the posterior mean and posterior mode must be obtained numerically
  - Let's focus on the mean

Bayesian Computation

- In general, we can approximate the mean of a distribution by averaging

Bayesian Computation

- In general, we can approximate the mean of a distribution by averaging
  - Given a sample, the average is a good approximation to the mean of the underlying distribution

Bayesian Computation

- In general, we can approximate the mean of a distribution by averaging
  - Given a sample, the average is a good approximation to the mean of the underlying distribution
- If we can generate a sample from the posterior, we can estimate the posterior mean

Bayesian Computation

- In general, we can approximate the mean of a distribution by averaging
  - Given a sample, the average is a good approximation to the mean of the underlying distribution
- If we can generate a sample from the posterior, we can estimate the posterior mean
- Bayesian computation is about efficiently generating a sample from the posterior distribution

Bayesian Computation

- In general, we can approximate the mean of a distribution by averaging
  - Given a sample, the average is a good approximation to the mean of the underlying distribution
- If we can generate a sample from the posterior, we can estimate the posterior mean
- Bayesian computation is about efficiently generating a sample from the posterior distribution
  - Often very computationally intensive

Bayesian Computation

- In general, we can approximate the mean of a distribution by averaging
  - Given a sample, the average is a good approximation to the mean of the underlying distribution
- If we can generate a sample from the posterior, we can estimate the posterior mean
- Bayesian computation is about efficiently generating a sample from the posterior distribution
  - Often very computationally intensive
  - Many tricks to improve performance

Bayesian Computation

- In general, we can approximate the mean of a distribution by averaging
  - Given a sample, the average is a good approximation to the mean of the underlying distribution
- If we can generate a sample from the posterior, we can estimate the posterior mean
- Bayesian computation is about efficiently generating a sample from the posterior distribution
  - Often very computationally intensive
  - Many tricks to improve performance
  - Often depends on the structure of the problem

Bayesian Computation

- Algorithms include:

Bayesian Computation

- Algorithms include:
  - Gibbs sampling

Bayesian Computation

- Algorithms include:
    - Gibbs sampling
    - Metropolis-Hastings

Bayesian Computation

- Algorithms include:
  - Gibbs sampling
  - Metropolis-Hastings
  - Approximate Bayesian Computation (ABC)

Bayesian Computation

- Algorithms include:
  - Gibbs sampling
  - Metropolis-Hastings
  - Approximate Bayesian Computation (ABC)
- There are also some analytic tools:

Bayesian Computation

- Algorithms include:
  - Gibbs sampling
  - Metropolis-Hastings
  - Approximate Bayesian Computation (ABC)
- There are also some analytic tools:
  - Laplace approximation

Bayesian Computation

- Algorithms include:
  - Gibbs sampling
  - Metropolis-Hastings
  - Approximate Bayesian Computation (ABC)
- There are also some analytic tools:
  - Laplace approximation
  - Variational Bayes

Bayesian Computation

- Algorithms include:
  - Gibbs sampling
  - Metropolis-Hastings
  - Approximate Bayesian Computation (ABC)
- There are also some analytic tools:
  - Laplace approximation
  - Variational Bayes
- Many, many more of both